Author/s:
Ramani, Rishi Sanjay

Title:
Machine learning and fluroescence in vivo confocal microscopy for the early detection of oral potentially malignant disorders and oral cancer

Date:
2025

Persistent Link:
https://hdl.handle.net/11343/361370

MACHINE LEARNING AND FLUORESCENCE IN VIVO CONFOCAL MICROSCOPY FOR THE EARLY DETECTION OF ORAL POTENTIALLY MALIGNANT DISORDERS AND ORAL CANCER

Rishi Sanjay Ramani

BDS, MDent, PCIP

ORCID: 0000-0002-3924-2133

Doctor of Philosophy

October 2025

Melbourne Dental School

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne

Submitted in total fulfilment for the degree of Doctor of Philosophy

Principal Supervisor

Prof. Michael McCullough

BDSc MDSc PhD FRACDS (Oral Med) FOMAA FPFA FICD

Oral Medicine Specialist & Professor

Melbourne Dental School, The University of Melbourne

Co-supervisors

Dr. Tami Yap

BDsc (Hons) DCD PhD FRACDS FOMAA FPFA

Oral Medicine Specialist & Senior Lecturer

Lead, Innovation and Engagement

Melbourne Dental School, The University of Melbourne

Dr. Lachlan Whitehead

BA BSc (Hons) PhD

Senior Research Officer

Advanced Technology and Biology

Walter and Eliza Hall Institute

Advisory Committee Chair

A/Prof. Catherine Butler

BSc (Hons) PhD

Principal Research Fellow

Melbourne Dental School, The University of Melbourne

ABSTRACT

Purpose: To utilise machine learning and fluorescence in vivo confocal microscopy for the early detection of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC).

Materials and Methods: In vivo confocal micrographs were captured in 59 patients at the Royal Dental Hospital Melbourne, Australia using InVivage® (Optiscan Imaging, Australia) with acriflavine (0.1%) and fluorescein (0.1%) as contrast agents. A systematic review was conducted to assess the use of confocal microscopy for the detection of oral cancer and oral potentially malignant disorders.

A convolutional neural network (CNN) model of the Inception_v3 architecture capable of filtering the confocal micrographs for diagnostic quality images was developed. The dataset for each contrast agent was divided into 4 diagnostic categories using histopathology as the reference: no dysplasia, lichenoid (chronic inflammation with lichenoid features), low-risk (low grade OED), and high-risk (high-grade OED & OSCC). The first diagnostic triage study involved qualitative features of cells and nuclei as identified by oral medicine specialists as inputs for diagnostic triage machine learning models.

The next study involved quantitative feature extraction of shape and pixel intensity measurements of cell nuclei as inputs for diagnostic triage machine learning models using StarDist, ImageJ, and Python. The next study involved development of Inception_v3 CNNs using transfer learning, cross validation and hyperparameter optimisation for diagnostic triage of micrographs.

For the last study in vivo, confocal micrographs were collected from 61 mice using the ViewnVivo (Optiscan Imaging, Australia) with acriflavine (0.1%) staining at the Walter and Eliza Hall Institute, Melbourne, Australia. CNNs were developed using the same method as the human dataset study to identify OED & OSCC.

Results: The systematic review on of confocal microscopy in oral cancer research identified both fluorescence and reflectance confocal microscope analysis along with qualitative and quantitative analysis for detection of carcinoma and dysplasia.

A total of 9168 confocal micrographs were captured across 59 patients. The quality filtering CNN study produced a model with accuracy of 89.5%. This CNN was used on the entire acriflavine and fluorescein image datasets to filter all available images which resulted in 1983 diagnostic quality images.

The qualitative human identified features and quantitative nuclei measurements showed significant relationships to the diagnostic criteria with the high-risk nuclei being brighter and larger than other nuclei. The top performing qualitative approach model

identifying 50% and 17% of low risk and high-risk lesions respectively. The quantitative approach model identified only 42.9% and 55.6% of low-risk and high-risk cases respectively.

The top performing diagnostic triage CNN identified 71% of low-risk cases and all of the high-risk cases correctly. The murine dataset diagnostic triage CNNs identified about 68.8% of low-risk cases and 75.6% high-risk cases.

Conclusions: The literature highlights heterogeneity in diagnostic criteria and the impact of assessor subjectivity in confocal microscopy aided oral cancer detection. While both qualitative and quantitative features on confocal micrographs showed significant relationships with diagnostic categories, the diagnostic triage CNN approach produced the best results in identifying low-risk and high-risk cases. The combined use of quality filtering and diagnostic CNN approaches can potentially speed up real-time precise diagnostic triage while maintaining imaging quality standards.

DECLARATION

This is to certify that:

(1) this thesis comprises only my original work towards the PhD except where indicated in the preface.

(2) due acknowledgement has been made in the text to all other material used.

(3) The thesis is fewer than 100,000 words in length, exclusive of tables, figures, bibliographies, and appendices.

(4) The thesis comprises 100% dissertation as agreed by the advisory committee at confirmation.

Rishi Sanjay Ramani

27st October 2025

PREFACE

This dissertation documents a series of studies based on the diagnosis of oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC) with the help of machine learning (ML) and fluorescence in vivo captured confocal microscopy images. The work presented in the following chapters, represents original research that I have conducted under the supervision and guidance of my supervisory panel.

The thesis consists of nine chapters, including one literature review, one systematic review, one comprehensive methodology chapter across all the work done, five chapters describing once individual study each and a discussion chapter.

The systematic review detailed in Chapter 2 was published in the journal *Oral Diseases* on 28th June 2022. The fluorescence in vivo confocal microscopy imaging protocol described in Chapter 3 (Methodology) was published in the journal *Frontiers of Oncology* on 3rd July 2023. Portions of the studies described in Chapter 4 (Quality filtering using machine learning) and Chapter 7 (Diagnostic triage using convolutional neural networks) was published in the journal *Scientific Reports* on 20th January 2025. The protocol for fluorescence in vivo confocal microscopy imaging in mice that was conducted at the Walter and Elize Hall Institute (WEHI) and described in Chapter 3 (Methodology) was published in the journal *Methods* on 24th April 2025.

Machine learning models and CNNs are incredibly complex mathematical entities. CNNs have millions of parameters. Thus, the entire complexity of these models was neither described nor discussed in this work. This is due to the fact that the scope of this work is limited to modifying and utilising these models for diagnostic triage and describing. Discussing the intricate details of each and every one of the millions of parameters and multitude of mathematical functions these models have is outside the scope of this work.

I hereby affirm that this thesis contains no material that has been submitted or accepted for the award of any other degree at any university or other tertiary institution. To the best of my knowledge and belief, this thesis does not include any material previously written or published by another individual, except where due reference has been appropriately cited in the text.

During my candidature I completed and graduated from the Professional Certificate of Innovation Practice (PCIP) at the University of Melbourne co-ordinated by the Innovation Practice Program (IPP) team in 2024. This six-month program focused on management, coaching and leadership skills required to guide innovation projects. Additionally, I completed the Translating Research at Melbourne (TRAMx) Bootcamp in 2023 which was an immersive bootcamp exploring research impact and entrepreneurship frameworks. I also participated in the Innovation By Design (IBD)

workshop within Professor Dougie Boyle's team for the design of large scale digital dental data infrastructure at the University of Melbourne.

Throughout my candidature I have been a member several professional associations including the International Association for Dental Research (IADR) and the Australasian Institute of Digital Health (AIDH). I have been an active member and project co-lead at the World Health Organisation (WHO) - International Telecommunication Union (ITU) - World Intellectual Property Organisation (WIPO) - Global initiative on AI 4 Health Topic Group - Oral Health since December 2023. Additionally, I have served as the Associate Editor for Bio-image analysis and Artificial intelligence at the Journal of Oral Pathology and Medicine since December 2023.

At the Melbourne Dental School (MDS), the University of Melbourne, I have been an active member of the Oral Medicine and Oral Cancer (OMOC) research group throughout my candidature. Additionally, I am a founding member of the Data, Dentistry and Artificial Intelligence (D2AI) research group at the MDS. During the past few months of my candidature, I was involved with development and delivery of the Digital Health curriculum for second year dental students at the MDS. My position as a research assistant at the MDS within oral medicine and dental data science began in February 2025 during the last few months of my candidature. Despite all my memberships and appointments, I have remained committed to my doctoral research project and pushed the boundaries of AI research in oral medicine.

I confirm that all work presented in this thesis was conducted during my PhD candidature. Additionally, I attest that no third-party editorial assistance was provided in the preparation of this thesis.

PUBLICATIONS AND PRESENTATIONS

FIRST-AUTHOR PUBLICATIONS ARISING FROM THIS THESIS

1. **Ramani R.S.**, Tan I., Bussau L., O'Reilly L.A., Silke J., Angel C., Celentano A., Whitehead L., McCullough M., Yap T. (2025) Convolutional neural networks for accurate real-time diagnosis of oral epithelial dysplasia and oral squamous cell carcinoma using high-resolution in vivo confocal microscopy. *Scientific Reports* 15, 2555 https://doi.org/10.1038/s41598-025-86400-5

2. **Ramani, R. S.**, Tan, I., Bussau, L., Angel, C. M., McCullough, M., & Yap, T. (2022). Confocal microscopy in oral cancer and oral potentially malignant disorders: A systematic review. *Oral Diseases*, 00,1-13. https://doi.org/10.1111/odi.14291

CO-AUTHORED PUBLICATIONS RELATING TO WORK IN THIS THESIS

1. Celentano A., Rickard J.A., Low J., Silke N., Mohammed A.I., Moslemi E., **Ramani R.S.**, Franca P.D., Reiner T., McCullough M.J., Yap T., Silke, J. & O'Reilly, L. A. (2025). Enabling high-resolution diagnostic oral confocal laser endomicroscopy in mice. Methods 239, 169-181 https://doi.org/10.1016/j.ymeth.2025.04.015
   I contributed to validation of the imaging protocol and reviewing the manuscript.

2. Yap T., Tan I., **Ramani R.S.**, Bhatia N., Demetrio de Souza Franca P., Angel C., Moore C., Reiner T., Bussau L., McCullough M.J. (2023). Acquisition and annotation in high resolution in vivo digital biopsy by confocal microscopy for diagnosis in oral precancer and cancer. Front. Oncol. 13:1209261. https://doi.org/10.3389/fonc.2023.1209261
   I contributed to validation of the imaging protocol and reviewing the manuscript.

CO-AUTHORED PUBLICATIONS ARISING DURING, BUT NOT RELATED TO, THIS THESIS

1. Kaur G., Yap T., **Ramani R.S.**, McCullough M., Singh A., (2024) Assessing bias in the casual role of HPV in oral cancer: A systematic review and meta-analysis. Oral Diseases, 00,1-9. https://doi.org/10.1111/odi.15062
   I assisted in the article screening stages and manuscript review.

2. **Ramani R. S.** (2024). Revolutionizing oral pathology and medicine: The artificial intelligence advantage. J Oral Pathol Med 2024;1-3. https://doi.org/10.1111/jop.13534
   I authored this editorial for the Journal of Oral Pathology and Medicine as the Associate Editor of Bio-image analysis and Artificial Intelligence.

3. Kaunein N., **Ramani R.S.**, Koo K., Moore C., Celentano A., McCullough M., Yap T. (2021). A Systematic Review of MicroRNA Signatures Associated with the Progression of Leukoplakia with and without Epithelial Dysplasia. Biomolecules 2021, 11, 1879. https://doi.org/10.3390/biom11121879
   I assisted in the article screening stages and manuscript review.

PRESENTATIONS AND POSTERS RELEVANT TO THIS THESIS

1. **Ramani, R.S.**, Tan I., Bussau L., Whitehead L., McCullough M., & Yap T. Machine Learning And Digital Biopsies For Oral Cancer Detection. Poster presented at the 103rd General Session and Exhibition of the International Association for Dental, Oral, and Craniofacial Research (IADR) in Barcelona, Spain on 25-27th June 2025.
2. **Ramani, R.S.**, O'Reilly, L.A., Whitehead, L., Low J., Bussau, L., Silke J., Celentano A., McCullough M., Yap T. High resolution confocal microscopy with deep learning for accurate real-time detection of graded oral epithelial dysplasia in a murine model of oral carcinogenesis. Poster presented at the 2nd Global Oral Cancer Forum (GOCF) at Kuala Lumpur, Malaysia on 25th-26th May 2024. This poster won the 1st prize at the poster competition held during this conference.
3. **Ramani R.S.**, Tan, I., Bussau, L., Whitehead L., McCullough, M.J., Yap, T. Fluorescence in vivo confocal laser endomicroscopy with deep learning for the early detection of oral potentially malignant disorders and oral cancer. Poster presented at the 4th National Light Microscopy Australia (LMA) conference held in Melbourne, Australia from 5th to 8th March 2024.
4. **Ramani R.S.**, Tan, I., Bussau, L., Whitehead L., McCullough, M.J., Yap, T. Deep Learning for Real-Time Analysis of In Vivo Captured Confocal Micrographs for Diagnostic Assessment of Oral Potentially Malignant Disorders and Oral Cancer. Oral presentation at the 9thWorld Congress of the International

Academy of Oral Oncology (IAOO) held in Incheon, Korea from 1st to 4th November 2023.

5. **Ramani R.S.**, Tan, I., Bussau, L., Whitehead L., McCullough, M.J., Yap, T. Rapid artificial intelligence assistance and in vivo captured micrography in oral cancer and potentially malignant disorders. Poster presented at European Association of Oral Medicine (EAOM) 16th Biennial conference held in London, United Kingdom from 29th & 30th September 2023. This poster won the prize for best poster at the conference.
6. **Ramani R.S.**, Tan, I., Bussau, L., Whitehead L., McCullough, M.J., Yap, T. Artificial Intelligence And In Vivo Confocal Microendoscopy In Oral Cancer. Poster presented at International Association for Dental Research (IADR) ANZ conference held in Sydney, Australia from 27th to 29th September 2023.

PRESENTATION OCCURING DURING BUT NOT RELATED TO THIS THESIS

1. **Ramani R.S.,** Jones B., Chaurasia A., & Tichy A. Knowledge and awareness of AI validation metrics in dental image analysis among dental AI researchers and clinicians. Oral presentation at the meeting of Topic Group Dental Diagnostics and Digital Dentistry, which is a part of the WHO-ITU-WIPO Global Initiative AI for Health at the Ludwig Maximillian University Hospital, Munich, Germany on 2nd July 2025.

AWARDS, SCHOLARSHIPS AND PRIZES

1. **Ramani R. S.** Oral Medicine Academy of Australasia Award (2023)
2. **Ramani R. S.** The Robert and Gillian Cook Family Award (2022-2025)
3. **Ramani R. S.** Melbourne Research Scholarship (Fee offset) (2022-2026)
4. **Ramani R. S.** Best poster award at the Global Oral Cancer Forum conference in Kuala Lumpur, Malaysia (2024)
5. **Ramani R. S.** Winner of the Colgate Hatton Poster competition Senior Category at the Melbourne Dental School (2025)
6. **Ramani R.S.** Runner up of the ANZ division Colgate Hatton Poster competition Senior Category for the International Association of Dental Research (IADR) (2025)

GRANTS AWARDED FOR WORK ARISING FROM THIS THESIS

1. **Ramani R.S.**, Yap T., Celentano A., McCullough M. (2023) Deep learning convolutional neural networks in in vivo micrographical diagnosis of oral cancer

and oral potentially malignant disorders. Australian Dental Research Foundation (ADRF) - $10,125

2. **Ramani R.S.**, Yap T., McCullough M. (2022-2025) Machine learning image analysis in confocal microscopy assisted diagnosis of oral potentially malignant disorders and oral cancer. Melbourne Dental School (MDS) Internal Grant - $1600

GRANTS AWARDED DURING BUT NOT RELATED TO THIS THESIS

1. Yap T., Unnithan R., Widdicombe B., Chau E., **Ramani R.S.**, McCullough M. (2024-2025) Development of a novel device “ThermOralCam” for the early detection of oral cancer. Graeme Clark Institute for Biomedical Engineering Proof-of-Concept Award - $50,000
2. Yap T., McCullough M., **Ramani R.S.**, Schwendicke F., Schwarzler J. (2024) Integrated AI Assistance in Community Oral Cancer Screening. Department of Medicine Dentistry and Health Sciences (MDHS) Innovation Seed Grant - $ 40,000

ACKNOWLEDGEMENTS

Our existence might have no meaning and anything we do or achieve could very well be pointless.

However, the love and kindness we receive from our fellow humans has the power to make us hitchhike all over the galaxy and back on less than thirty Altairian dollars a day. I am lucky to have received such love, and I am using this platform of my life's defining work to pay my gratitude to those who supplied it.

My parents Meenakshi and Sanjay have been life defining role models without whom I would not have been capable of my journey. They set me up for success right from the beginning with their constant and ever-dependable support. Their accomplishments and vast experiences provided an anchor for my perspective on the world. My father's ability to overcome any obstacle with wit and confidence without even acknowledging the concept of a 'comfort zone' has emboldened me to push my own boundaries. My mother's calm strength and unwavering moral compass regardless of the challenge she is faced with provided me a powerful foundation to build my identity.

My grandmother Sulochana, who had a major role in raising me, always possessed a resolute determination to never fall short in anything she tried which fuelled my competitive spirit. The cooking skills she has instilled in me have been my shield against home-sickness. Although my grandfather Kishore left us early, his discipline and strong sense of integrity as echoed throughout my life decisions. My grandmother Sushila has been an inspiration of hard work with her love language of delicious food always reminded me how good life is. My grandfather Natwarlal was a brilliant, kind, and well-read man whose love for storytelling planted the seeds of curiosity in my mind.

My extended family have each had their own unique imprint on my life. Bharti's dedication to science and academia inspired me to pursue it. Paresh's passion for experimentation guided me throughout my journey. Vaishali's love for learning infected me. Shilpa's resilience taught me about endurance and hope. Pankaj's humility showed me the power of good leadership. Jyoti's adaptability to unfamiliar environments inspired malleability. Ishan's generosity reminded me of the quiet power of care. Drishti's warmth reminded me of the joy in companionship. Shibanni's adventurous spirit taught me that curiosity is its own reward. Sonjana's creativity and joyful spirit reminded me of the lightness in learning. Rushabh's humour, endless patience, and silent competitiveness reminded me that ambition and warmth can coexist.

My in-laws Amruta and Rajan have always shown their support of my academic ventures with patience and kindness. My brother-in-law Raj's calm spirit and quiet intelligence taught me the value of patience in thinking and living.

My close friends have been my constants on this journey. Aditya's companionship and willingness to listen gifted me the space to think and grow. Rajat's sharp mind and love for strategy games taught me to think several moves ahead. Aamir's willingness to engage in endless reflection gave me clarity in uncertainty.

I am deeply grateful to my colleagues & friends - Ivy, Huda, Suhaib, Satutya, Andrew, Ayu, Brian, Caroline, and Hamza - whose thoughtful feedback and the countless conversations enriched my research and made the journey all the more enjoyable. Antonio's ever watchful and supportive guidance shaped my journey. Chau's wisdom and support always ensured I had the facilities I needed. Yeganeh's cheerfulness and infectious positivity showed me that friendship makes every challenge easier to bear.

Nadia's friendship and timely advice, having walked this path just ahead of me, carried me through the hardest stretches of my PhD. Bree's kindness and understanding made my PhD very enjoyable, and our continuing collaborations remind me how enduring bonds enrich both life and research.

I owe deep gratitude to my supervisors for their mentorship, constructive criticism, and constant belief in my work. Lachlan's expertise in bio-image analysis guided my work with precision, and his relaxed and accommodating approach made his brilliant insights all the more impactful. Catherine's nurturing influence, unwavering support, and gentle wisdom gave me the belief that my work was truly special. Vincent taking me under his wing when I needed it the most is a kindness I will always cherish. I am especially grateful to Joanne whose invaluable mentorship and inclusive approach gave me the platform to find my identity as a researcher. The kindness Alison showed me by giving me a chance in the oral medicine world is something I can never forget to appreciate.

And most importantly Michael and Tami, who are two of my favourite people in the world. I am eternally grateful for this opportunity to learn from them.

Michael has been an absolute inspiration throughout the time I have known him. His wealth of knowledge and passion for science never failed to amaze. Our one-on-one meetings always doubled the amount of time planned because of our long conversations about life, the universe and everything. His big picture perspective always helped me reflect on the impact of my project rather than getting stuck in the technical weeds. I genuinely feel he went above and beyond his role as a PhD supervisor simply based on how much time he spent staring at a computer screen and tracing circles on it. He is an absolute legend in his field yet talking to him always made me feel heard and respected.

Tami is simply brilliant. She is exceptionally good at everything she does and bends the rules of space and time to accomplish so many amazing feats so quickly. There were times where I felt like I might not be good enough to be her student. I've been pretty hard on myself at times during my doctoral studies, but Tami always seemed to sense it and somehow told me exactly what I needed to hear whether it was encouragement or a reality check. Her amazing philosophy of inclusion and directness with collaborations

gave me opportunities to connect with some brilliant people who have all made my research skills better and in turn improved my self-confidence by magnitudes. She is the kind of mentor I aspire to be.

Carl Sagan, Frank Herbert, Issac Asimov, Andy Weir, Christopher Nolan, and Jake Peralta among other visionaries have showed me the beauty in our world and theirs. Douglas Adams' playful wisdom made me realise that the answer may be 42, but the joy lies in the questions. Terry Pratchett's imagination and satire taught me that vigilance, stubbornness, and a little bit of cynicism can see you through even the longest nights of a PhD. The universes and stories of Skyrim, Elden ring, Warhammer, Dark Souls, and Geralt of Rivia reminded me that epic struggles are always worth undertaking despite going up against overwhelming odds. Essendon Bombers' endurance and Seattle Seahawks' fighting spirit inspired me to keep faith in the long game.

My puppy Remi has been a much-needed source of unconditional love (if you don't count food and cuddles). Her boundless energy and pure affection dissolve my anxieties and shields me from the weight of the world.

Finally, the love of my life, and my safest haven, my ever-loving wife, Ruhi. Words are not capable of describing what she means to me. She is the smartest, kindest, most beautiful person in my world. I would literally not be here without her. She has had to be incredibly patient with me these past few years while tolerating my mad ramblings, being my counsellor, and challenging me in the best of ways. Her own accomplishments continue to inspire me and knowing that I can always rely on her keeps me going. I am exceptionally lucky to share my home and life with her.

"Real stupidity beats artificial intelligence every time."

- Sir Terry Pratchett, *Hogfather* (1996)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PROTOCOLS

LIST OF CODE STRUCTURE OUTLINES

ABBREVIATIONS

2D	2-dimensional
AC	Actinic cheilitis
AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application programming interface
ASIR	Age-standardised incidence rates
ASMR	Age-standardised mortality rates
AUROC	Area under the receiver operator characteristic curve
CLE	Confocal laser endomicroscopy
CNN	Convolutional neural networks
CSV	Comma separated values
CT	computed tomography
DL	Deep learning
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
FCM	Fluorescence confocal microscopy
FDA	Food and Drug Administration
FN	False negatives
FOV	Field-of-view
FP	False positives
GI	Gastrointestinal tract
GPU	Graphics processing unit
H&E	Haematoxylin & eosin
IARC	International Agency for Research on Cancer
ID	Identifier
IHC	Immunohistochemistry
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoU	Intersection over Union
ITU	International Telecommunication Union
LOCI	Laboratory for Optical and Computational Instrumentation
LR	Logistic regression
MDS	Melbourne Dental School
ML	Machine learning
MMS	Mean matched score
MRI	magnetic resonance imaging

MTS	Mean true score
NMS	Non-maximum suppression
NPV	Negative predictive value
OCR	Optical Character Recognition
OCT	Optical coherence tomography
OED	Oral epithelial dysplasia
OLL	Oral lichenoid lesions
OLP	Oral lichen planus
OMOC	Oral medicine and oral cancer group
OPMD	Oral potentially malignant disorders
OSCC	Oral Squamous Cell Carcinoma
OSMF	Oral submucous fibrosis
PAI	Photo-acoustic imaging
PET	positron emission tomography
PPV	Positive predictive value
PQ	Panoptic quality
PSC	Point scanning confocal
QMR	Quality Micrograph Refiner
RAM	Random access memory
RCM	Reflectance confocal microscope
RDHM	Royal Dental Hospital Melbourne
ReLU	Rectified Linear Unit
RF	Random forest
RNA	Ribonucleic acid
ROC	Receiver operator characteristic
ROI	Region of interest
S.d.	Standard deviation
SGD	stochastic gradient descent
SVM	Support vector machines
TIFF	Tagged image file format
TN	True negatives
TP	True positives
TPU	Tensor processing unit
WCSS	Within cluster sum of squares
WHO	World Health Organization
WIPO	World Intellectual Property Organization
XGB	XGBoost
ZeroCostDL4Mic	Zero Cost Deep Learning for Microscopy

1. LITERATURE REVIEW

1.1. Introduction

Oral cancer is a serious and often life-threatening condition that affects millions of individuals globally. Early detection is crucial for improving survival rates and reducing the need for aggressive treatment. Cancer prevention encompasses primary, secondary, and tertiary prevention. Primary prevention consists of actions that lower the risk of developing cancer and tertiary prevention focuses on actions which can reduce the impact of cancer once it is already established. The focus of this work is on secondary prevention, which involves the use of methods for the detection of asymptomatic or early symptomatic precancerous conditions or cancer at a stage when lesions can be more easily treated (IARC, 2023).

Traditional diagnostic methods, such as surgical biopsies, are often invasive, time-consuming, and dependent on subjective evaluation. Recent breakthroughs in digital microscopy and imaging technology are transforming this landscape. These innovations offer non-invasive, highly accurate, and real-time diagnostic solutions that enable the identification of cancerous and precancerous lesions in the oral cavity with potential unprecedented precision. These technologies may not only enhance the capabilities of clinicians but also improve patient experiences. The integration of artificial intelligence (AI) further amplifies the diagnostic power of these methods, ensuring earlier and more accurate detection.

The romantic vision of machines that can think by themselves in their service to humanity has been a core theme in science fiction. Visionaries such as Issac Asimov, Philip K. Dick, and Douglas Adams explored these ideas while drawing caution to the darker elements that could disrupt this utopian dream. The innovation of AI has irreversibly changed our world. This movement shows promise of a future with enhanced prevention, timely intervention, and improved outcomes in healthcare. However, a lack of understanding of AI could lead to pitfalls deeper than we ever imagined and cause irreparable damage to humanity.

1.2. Anatomy of the mouth

The mouth or oral cavity is the entrance to the gastrointestinal tract, which is bounded anteriorly by the lips, superiorly by the palate, laterally by the cheeks, inferiorly by the mylohyoid muscle floor and posteriorly by the faucial pillars (Standring, 2020).

The oral soft tissue occupying this space is covered by a mucous membrane comprising of stratified squamous epithelium. This oral mucosa can be keratinised (masticatory mucosa), non-keratinised (lining mucosa), or specialised mucosa and has the ability to undergo regeneration (Nanci, 2017) (Figure 1.1.).

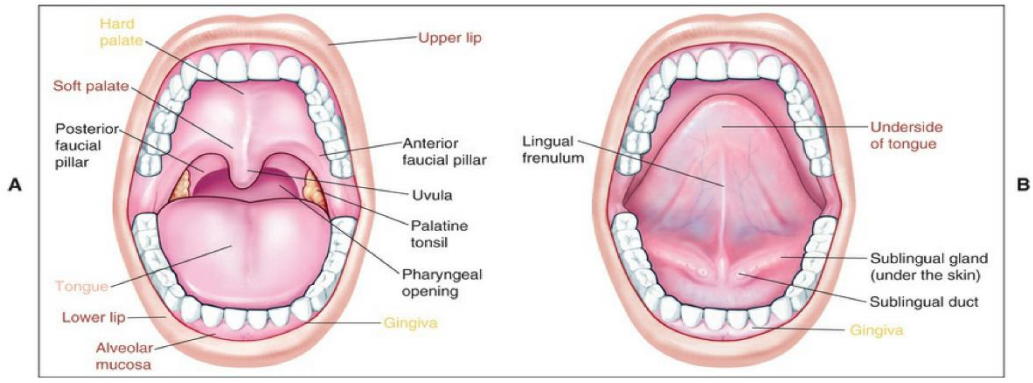


Figure 1.1. Components of oral anatomy (taken from Ten Cate's 9th edition: Nianci, 2017)

The intra-oral subsites covered by oral mucosa as described in the International Classification of Diseases for Oncology (ICD-O) are the lip, tongue, gingivae, floor of the mouth, hard palate, soft palate, and buccal mucosa. The lips form the external boundary of the mouth by surrounding the oral aperture. The lips contain a vermillion border also known as the lip line, which is a distinct boundary that separates the pink/red part of the lips and the surrounding skin and acts as a transition zone between the keratinised skin of the face and the non-keratinised epithelium of the lip. The tongue is a highly vascularised muscular organ. The important anatomical areas of the tongue are the keratinised anterior dorsum, lateral borders, and non-keratinised ventral (undersurface) surface. The gingiva is mucosa that surrounds the cervical margin of erupted teeth where the enamel ends, and the cementum of the tooth root begins. The floor of the mouth is a region of mucosa that separates the movable part of the tongue from the mylohyoid muscles. The hard palate consists of keratinised mucosa that forms the roof of the oral cavity. This mucosa contains transverse, irregular, and asymmetric ridges of mucosa located in the anterior third of the palate called palatal rugae. The soft palate is a mobile

flap that extends backward from the hard palate that is connected to the lateral walls of the oro-pharynx via the palatoglossal and palatopharyngeal arches which are together known as the pillars of fauces. Between these two arches lies the tonsillar fossa, which houses the palatine tonsils. The buccal mucosa is the inner lining of cheeks extending from the lips to the pterygomandibular raphe (Nanci, 2017; Standring, 2020; WHO, 2013).

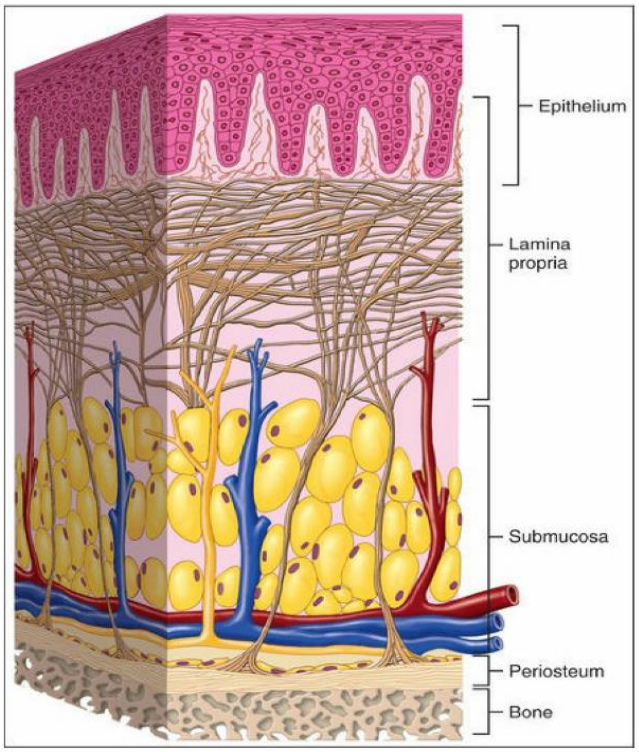


Figure 1.2. Main components of oral epithelium. Adapted from Ten Cates 9th edition, Nianci et al., 2017

The oral mucosa consists of two main components: the oral epithelium, a stratified squamous epithelium, and the underlying connective tissue known as the lamina propria (Figure 1.2.). The oral epithelium acts as a protective barrier between the deeper tissues and the oral environment. Its cells are tightly connected and arranged in distinct layers. The masticatory mucosa (found in areas like the gingiva, hard palate, and dorsal tongue) has a surface that is tough, inflexible, abrasion-resistant, and firmly attached to the lamina propria. This surface is covered by a layer of keratinized cells, produced through a process called keratinization. The innermost layer, or basal layer (stratum basale), consists of cuboidal or columnar cells adjacent to the basal lamina. Above this is the prickle cell layer (stratum spinosum), composed of elliptical or spherical cells connected at points called intercellular bridges or desmosomes. Together, the basal and prickle cell layers account for about half to two-thirds of the

epithelium's thickness. Above the prickle cell layer is the granular layer (stratum granulosum), which contains larger, flattened cells with keratohyalin granules that stain intensely with acidic dyes like hematoxylin. The outermost layer, the keratinised layer (stratum corneum), consists of flat, eosinophilic cells (called squames) that lack nuclei and stain bright pink with eosin (Nanci, 2017).

In contrast, the lining mucosa, found on the lips, buccal mucosa, alveolar mucosa, soft palate, underside of the tongue, and floor of the mouth, typically has a non-keratinised epithelium. While the basal and prickle cell layers in non-keratinised epithelium are similar to those in keratinised epithelium, the cells are slightly larger, and the intercellular bridges are less distinct. Non-keratinised epithelium lacks a granular layer, and the outer half is divided into the intermediate layer (stratum intermedium) and the superficial layer (stratum superficiale). The cells in the superficial layer retain their nuclei, which are often plump, and this layer does not stain as intensely with eosin as the surface of keratinised or parakeratinised epithelium (Nanci, 2017).

1.3. Oral cancer and oral potentially malignant disorders

Oral cancer is one of the most common types of cancer globally with an estimated 377,713 new cases identified worldwide in 2020 as reported by the International Agency for Research on Cancer (IARC) as a part of the World Health Organization (WHO) (IARC, 2023) (Figure 1.3.). In 2020, there were an estimated 177,757 deaths from oral cancer recorded globally with the age-standardised mortality rates (ASMR) of 2.8 per 100,000 males and 1 per 100,000 females (IARC, 2023). In 2023 the age-standardised incidence rates (ASIR) were 6.0 per 100,000 males and 2.3 per 100,000 females (IARC, 2023).

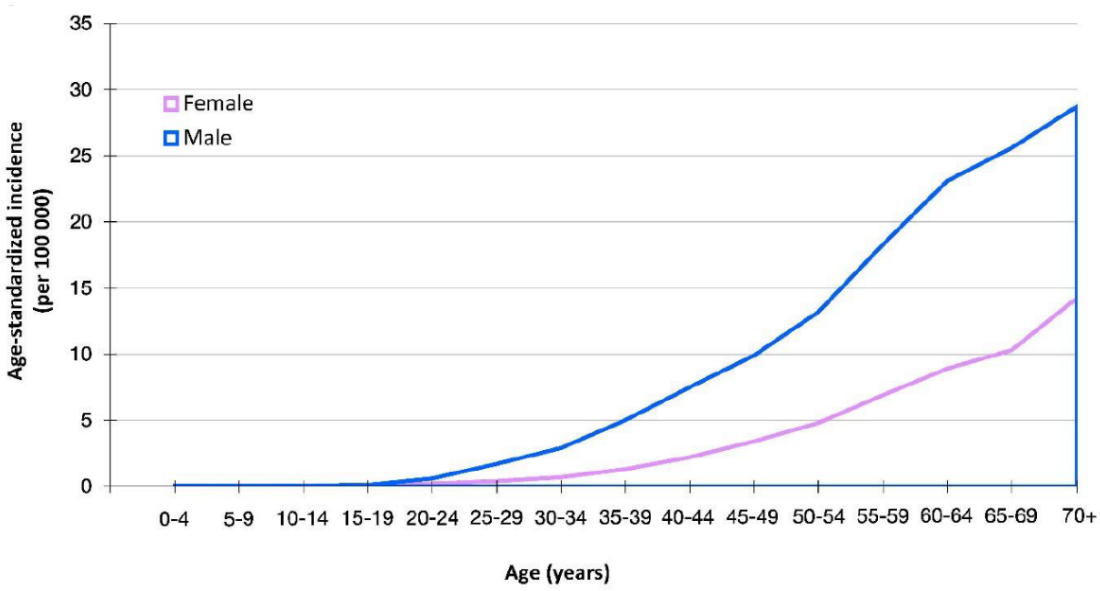


Figure 1.3. Age-specific incidence curves for oral cancer in 2020 from Ferlay et al. (2020) as depicted in IARC Handbook 19 (2023)

The incidence and mortality rates of oral cancer in 2020 were the highest in Melanesia and South Asia with ASIR and AMIR being consistently higher in men than women (Ferlay et al., 2020). Globally the projected increase in number of new oral cancer cases from 2020 to 2024 is projected to be 49.6% per year (Ferlay et al., 2020) (Figure 1.4.). The most common type of oral cancer is oral squamous cell carcinoma (OSCC) which constitutes over 90% of oral cancer cases. This data underscores the critical importance of early diagnosis in reducing the mortality rate associated with oral cancer within the population. OSCC is the focus of this work, and all mentions of 'oral cancer' in this dissertation beyond this point refer to OSCC, unless otherwise specified.

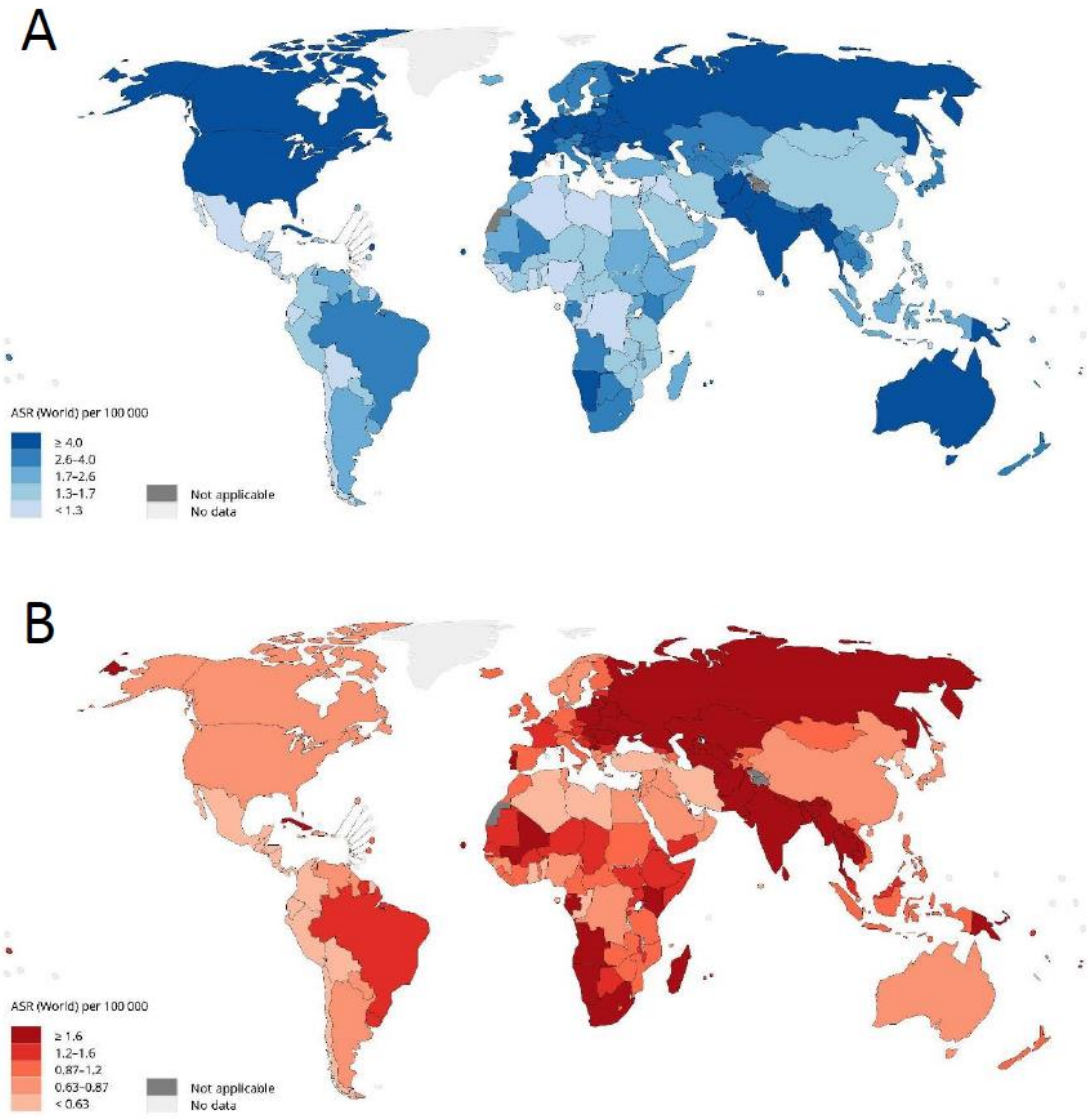


Figure 1.4. Global distribution of estimated age-standardised (world) (A) incidence rates and (B) mortality rates per 100,000 for oral cancer in 2020 as depicted in the IARC Handbook 19 (2023)

To effectively implement secondary prevention of oral cancer, it is essential to identify, at an early stage, precursor lesions that have a high likelihood of progressing to malignancy. The WHO working group defined oral potentially malignant disorders (OPMD) as "any oral mucosal abnormality that is associated with a statistically increased risk of developing oral cancer" (Warnakulasuriya, Kujan, Aguirre-Urizar, Bagan, González-Moles, Kerr, Lodi, Mello, Monteiro, Ogden, et al., 2021). Clinically these conditions present a wide range of features including colour variations, topographic changes, and variable size. They can be described as leukoplakia, erythroplakia, oral lichen planus,

oral lichenoid lesions, oral submucous fibrosis, actinic cheilitis, oral lupus erythematosus, dyskeratosis congenita, and graft versus host disease (Warnakulasuriya, 2020b). While these conditions pose an increased risk for cancer of the oral cavity only a small proportion of them progress to carcinoma during the life of the patient (Warnakulasuriya, Kujan, Aguirre-Urizar, Bagan, González-Moles, Kerr, Lodi, Mello, Monteiro, Ogden, et al., 2021).

The common clinical characteristic features noted by oral health professionals while diagnosing OPMDs or oral cancer are: site, clinical appearance, size, patient age, duration of lesion, sex, and habits (Speight, Khurram, & Kujan, 2018b). The screening methodology consists of identifying signs of abnormality, including changes in colour, texture, ulceration, and swelling. The most common colour associated with oral malignant or potentially malignant lesions is white with the second most common being red. Despite clinical appearance being somewhat unreliable, it has been used in the past as a predictor of malignant transformation. The commonly reported flat, white, homogenous plaque-like lesions show a low rate of malignant transformation (4-18%), regardless of the extent of dysplasia (Tsantoulis, Kastrinakis, Tourvas, Laskaris, & Gorgoulis, 2007). On other hand lesions which appear red or a combination of red and white have shown higher rates of malignant transformation (14-50%) (Tsantoulis et al., 2007).

The location of the lesion in the oral cavity could have an influence on its risk of transformation. However, these sites vary according to the specific habits of the patient. The lateral border of the tongue and floor of the mouth are found to be the most common sites for OPMDs and oral cancer (Speight, Khurram, & Kujan, 2018a). A study in Australia found that 40% of the lesions suspected to be dysplastic or malignant were found on the tongue and floor of the mouth (Dost, Lê Cao, Ford, Ades, & Farah, 2014). They also observed 30% of the lesions in this study to be on the buccal mucosa (Dost et al., 2014). However, these lesions were found to be less likely to be dysplastic (Dost et al., 2014). On the other hand, Holmstrup et al. (2006) followed up on 269 cases of potentially malignant lesions and found that the site of the lesions was not a significant prognostic factor. They found that the size of the lesions and whether they were homogenous or non-homogenous were significant more important in predicting malignant transformation (Holmstrup, Vedtofte, Reibel, & Stoltze, 2006).

Dost et al. (2014) further reported a retrospective audit of biopsy reports recorded at the Queensland Medical Laboratories (Australia) between 1995-2014. They reviewed reports of 383 lesions across the oral cavity. Out of all the lesions reviewed almost half (187 lesions) were observed on the tongue. This site showed the highest transformation rate of 7% (13 lesions transformed) among all the oral sites studied (Dost et al., 2014). There is controversy

regarding the various factors that determine the probability of malignant transformation of OPMDs. However, the importance of early stage diagnosis remains, with prognosis of a higher five-year survival rate reaching over 84% (Program, 2022).

Oral lichen planus (OLP) and oral lichenoid lesions (OLL) are chronic, non-infectious, and inflammatory lesions of the oral mucosa that are associated with specific external triggers or systemic conditions (Carrozzo, Porter, Mercadante, & Fedele, 2019; Lodi et al., 2005). OLP and OLL are recognised as distinct conditions and have different reported rates of malignant transformation in previous systematic reviews, ranging from 0.9-10.9% for OLP and 2.5%-3.2% for OLL (Aghbari et al., 2017; Fitzpatrick, Hirsch, & Gordon, 2014). However, they share several clinical, histopathological, and immunological features (Van der Meij & Van der Waal, 2003). These lesions have been considered to be nearly identical from a histopathology perspective (Isaac Van der Waal, 2009). Previous studies by Van der Meij et al. (2003) and Speight et al. (2018) have established that the presence of OED excludes the diagnosis of OLP, and those lesions are managed as dysplastic lesions (Speight et al., 2018b; Van der Meij, Mast, & van der Waal, 2007). The controversial nature of previously reported literature malignant transformation of OLP and OLL due to questionable application of diagnostic criteria has been scrutinised (Gonzalez-Moles, Scully, & Gil-Montoya, 2008). Celentano & Cirillo (2024) recently highlighted the issues with the categorisation of OPMDs based on variable rates of malignant potential (Celentano & Cirillo, 2024). The various opinions expressed in literature denote the difficulties in clearly categorising these lesions.

The early diagnosis of oral squamous cell carcinoma (OSCC) necessitates a combination of conventional expert oral examination and scalpel biopsy. Considering the diagnostic tools available presently, the gold standard in diagnosis of malignant lesions and OPMDs is incisional biopsy followed by a histological analysis (M. McCullough, G. Prasad, & C. Farah, 2010a). During histological examination, the architecture of the epithelium with distortions in cellular and nuclear structure are analysed. Abnormalities are referred to as oral epithelial dysplasia (OED) (Pindborg, Reichart, Smith, & Van der Waal, 2012). The WHO classifies oral epithelial dysplasia (OED) into mild, moderate, and severe dysplasia. Mild dysplasia is cytological or architectural alterations in the lower one-third of the epithelium. Moderate dysplasia relates to epithelial atypia which extends to the spinous epithelial layer at the middle third. Severe dysplasia is represented by epithelial atypia which extends throughout the entire thickness of the epithelium and is sometimes identified as carcinoma-in-situ (Reibel et al., 2017a). Another proposed method of dysplasia classification is the binary system which divides OEDs into low-grade and high-grade (Kujan et al., 2006). The low-grade classification is akin to the ‘mild dysplasia’ label from

the WHO system and the high-grade dysplasia lesions are a combination of the 'moderate' and 'severe dysplasia' lesions identified in the WHO system (Reibel et al., 2017a). Some of the features of dysplasia used to identify the grading of OEDs are displayed in Table 1.1.

Pathologists' interpretation of the criteria for dysplasia observed in any sample can be highly subjective and studies note a variation between observations recorded by different pathologists (Warnakulasuriya, 2020a). Kujan et al. (2007) attempted to understand this variation in professional opinion among pathologists. They found that pathologists often agreed upon observations of drop-shaped rete ridges, abnormal variation in cell shape, increase in mitotic figures, and increase in nuclear size. On the other hand, majority of the disagreements between pathologists arose regarding loss of basal cell polarity, hyperchromatism, and irregular epithelial stratification (Kujan et al., 2007). To add to the complexity of diagnosis, differential factors have the potential to cloud the results of histology.

Table 1.1. Features of oral epithelial dysplasia as presented by Reibel et al., 2017

Architectural features	**Cytologic features**
Irregular epithelial stratification	Abnormal variation in nuclear size (anisonucleosis)
Loss of basal cell polarity	Abnormal variation in nuclear shape (nuclear pleomorphism)
Drop-shaped rete ridges	Abnormal variation in cell size (anisocytosis)
Increased number of mitotic figures	Abnormal variation in cell shape (cellular pleomorphism)
Abnormally superficial mitoses	Increased nuclear/cytoplasmic ratio
Premature keratinisation in single cells (dyskeratosis)	Atypical mitotic figures
Keratin pearls within rete ridges	Increased number and size of nucleoli
Loss of epithelial cell cohesion	Nuclear hyperchromatism

Histopathological analysis has its drawbacks such as interpretation errors and difficulty in tracking lesion progression (Holmstrup et al., 2006). Another deterrent in the efficacy of histopathological assessment is the finding that the degree of dysplasia does not necessarily correlate with the potential for

malignant transformation (McCullough et al., 2010a). Invasive intraoral biopsies are commonly conducted under local anaesthesia in outpatient settings. However, these procedures often induce patient anxiety and lead to post-procedural oral pain. Moreover, histopathological findings derived from excised tissue samples are typically extrapolated to the remaining abnormal mucosa, which may not fully represent the pathology. Consequently, non-invasive diagnostic tools are being actively developed to provide cellular-level resolution, rapid assessment, and multisite analysis, extending beyond the capabilities of white-light visualization of the oral mucosa (Yap et al., 2023).

1.4. Confocal microscopy

Analysing biopsy specimens on a table-top microscope has been the preferred modality for the diagnosis of intra-oral lesions. The limitations to this procedure range from processing time, costs, patient morbidity, and potential sampling errors. Interpretation of these biopsy samples could be a subjective process. This led to development of real time in vivo imaging (Chu, 2010).

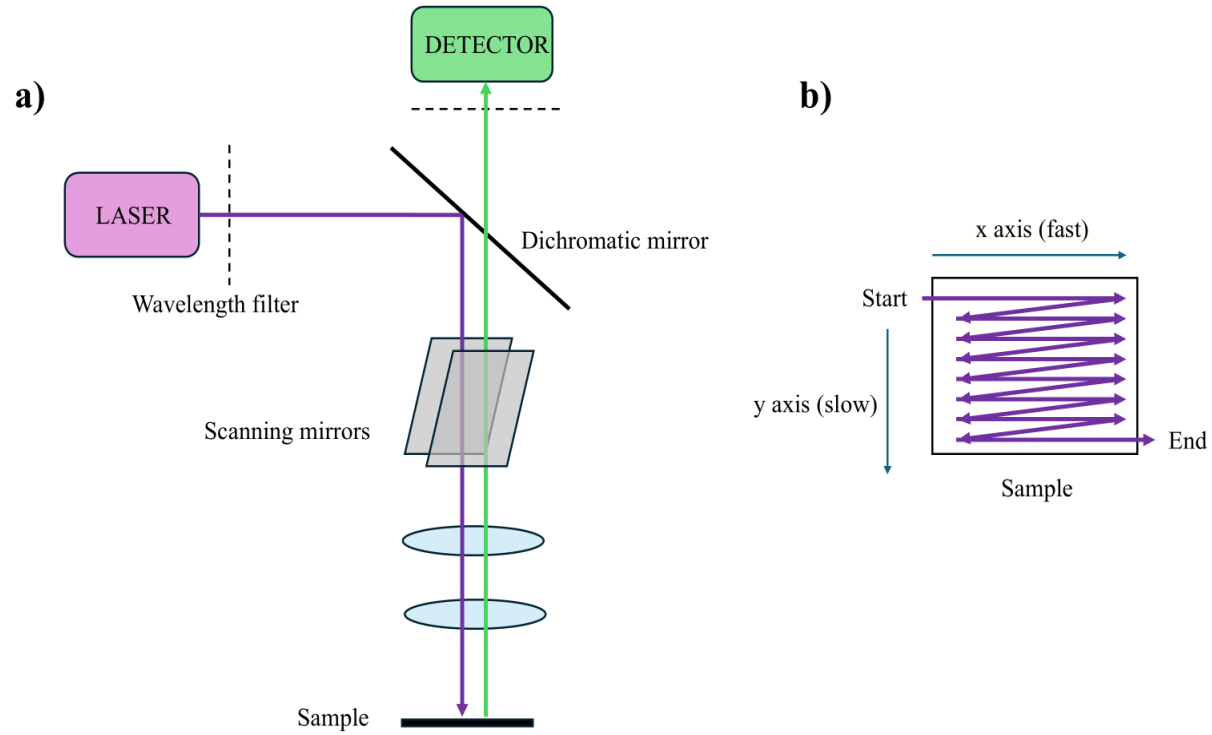


Figure 1.5. (a) Schematic diagram of a confocal laser scanning microscope, illustrating the key optical components: laser source, wavelength filter, dichromatic mirror, scanning mirrors, focusing lens, sample, emitted light path, and detector. This setup enables point-by-point illumination and detection for high-resolution imaging. (b) Raster scanning pattern used in confocal microscopy, demonstrating sequential laser scanning across the sample's x and y axes to construct a cross-sectional image.

An in vivo confocal microscope is one such device that has the potential to provide real time imaging of intra oral tissues with sub-cellular resolution and a reasonably high frame rate. In the 1950s, Marvin Minsky invented and patented the confocal microscope to visualize neural networks in 3D, a challenge for conventional microscopy. His design used aligned pinholes to focus on a single point, creating sharp, in-focus "pixels" by scanning across the sample in a raster pattern. The raster pattern in digital confocal microscopy refers to the systematic, line-by-line scanning of a specimen typically from left to right and top to bottom using a focused laser beam to construct a high-resolution image pixel by pixel (Figure 1.5.). Despite lacking modern technologies like lasers and

digital imaging, Minsky proved the concept, laying the groundwork for today's confocal laser scanning microscopes (Chu, 2010). Confocal microscopes are known for their sharp images owing to the placement of a 'pinhole' between the detector and the objective lens. The pinhole is used to filter light only coming from the focal plane of the lens, and reject any photons emitted from above and below despite those regions being illuminated due to the geometry of the beam paths. This provides the ability to filter in light from a single small source on the sample and reject the scattering rays from the surroundings that could distort the image (Chu, 2010).

While large tabletop microscopes have been incredibly effective for detailed imaging in labs, using this technology in medicine for real-time imaging brought a new challenge: making it portable. One way to achieve this was to create a hand-held version, something small and easy to use that could be taken directly to patients. This shift aimed to make imaging more accessible, allowing doctors to perform diagnostics at the bedside instead of requiring patients to be transported to specialised facilities. While designing this hand-held miniature version of this microscope, the core of the optical fibre assumes the role of the 'pinhole'. A major challenge in using this hand-held technology is the high performance demanded from miniature microscopes that includes a high resolution, fast frame rates and an adequate imagine depth for diagnosis (T. D. Wang, Mandella, Contag, & Kino, 2003).

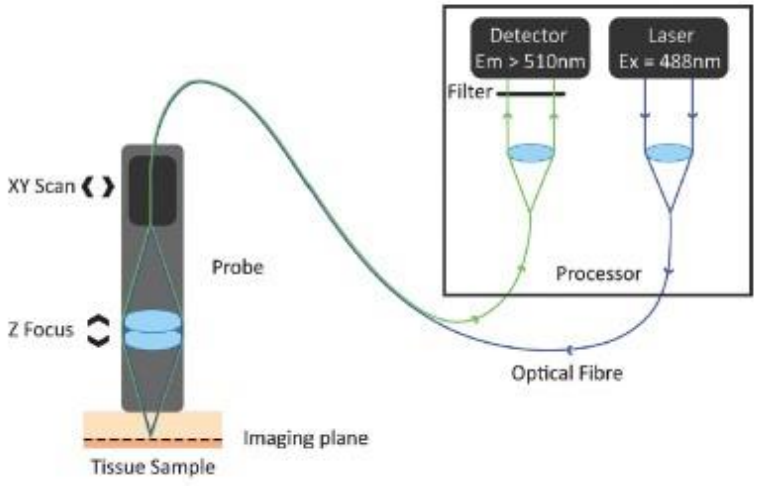


Figure 1.6. Working principle of a fibre optic laser confocal microscope. Adapted from Rangrez et al. (Part 1, 2021)

In the 1980s, Australian inventor Martin Harris discovered that the tip of an optical fibre could serve as both the illumination and detection pinholes in a confocal microscope (Figure 1.6.). This innovation simplified the design by eliminating the need for complex alignment. Along with Peter Delaney, Harris leveraged this breakthrough to enable the miniaturisation of confocal

microscopes, leading to the invention of confocal laser endomicroscopy (CLE) (Delaney, Harris, & King, 1994).

CLEs create an image by scanning a focused laser beam across a specimen in a raster pattern using two high-speed oscillating mirrors. One mirror scan along the x-axis, while the other shifts the beam incrementally along the y-axis (Figure 1.5.a). After each x-axis scan, the beam rapidly resets to the starting point (flyback) and moves to the next y-axis position, during which no image data is collected (Figure 1.5.b). This process excites a defined area in the specimen's focal plane. Fluorescence emitted by the specimen follows the same optical path as the excitation beam and is directed through a pinhole aperture to the detector. Unlike the moving excitation light, the emitted fluorescence remains stationary at the pinhole but fluctuates in intensity as the beam scans the sample. A photomultiplier converts this emission into an analogue electrical signal, which is digitized into pixels by an analogue-to-digital converter. The digital data is temporarily stored in the computer and displayed as an image. Importantly, the confocal image is reconstructed point by point from photon signals and does not exist as a real image that could be observable through a traditional microscope eyepiece.

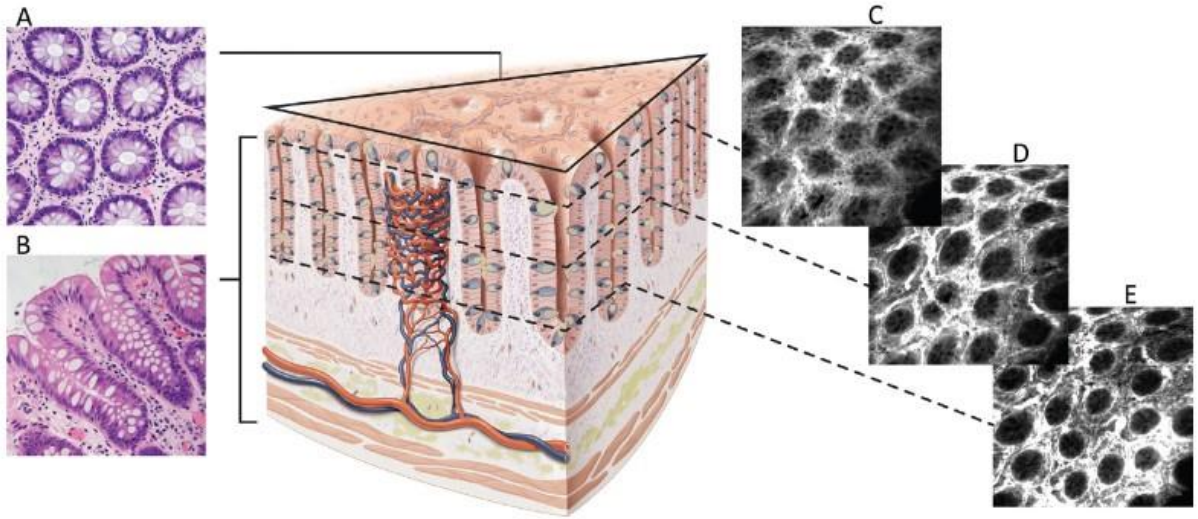


Figure 1.7. Difference in output from confocal microscopy compared to histopathology. Adapted from Rangrez et al. (Part 1, 2021)

This raster pattern point-scanning mechanism using a single fibre for both illumination delivery and incoming detection of fluorescence has superior resolution to the bundle fibre approach. Bundle fibres consist of a bundle core with each fibre capturing images allowing for rapid capture but resulting in an image with missing data between the bundles, thus limiting resolution ability (Rangrez, Bussau, Ifrit, & Delaney, 2021). The point scanning approach on the

other hand results in an image where each point is in focus, leaving no empty spots while also allowing for dynamic depth actuation. This enables capturing high-definition images at submicron resolutions with a large field-of-view (FOV) across different focal planes along a Z-axis (Figure 1.7.). Point Scanning Confocal (PSC) is a patented technology designed and developed by Optiscan and has been utilized in pre-cancerous conditions of the oesophagus, cervix, and colon (Canto et al., 2014; Hurlstone et al., 2008; J. Tan, Quinn, Pyman, Delaney, & McLaren, 2009). The technology has previously been utilised in studies using fluorescent dyes for tissue visualization of the cervix (J. Tan et al., 2009).

This PSC approach enables the capturing of highly defined megapixel images at submicron resolutions with a large field of view. This capability is combined with the ability to shift the focal plane of the handheld microscope lens along the Z-axis orientation of the tissue being imaged. In case of imaging the surface epithelium, this indicates that different layers of cells can be visualised along the depth of the tissue (Figure 1.7.). This miniature microscopy technology has been used to design the Fluorescence In Vivo Endomicroscopy (FIVE) instrument series by an Australian company called Optiscan Imaging (Victoria, Australia). These FIVE microscopes have the ability to optically section the tissue via 'digital biopsies' to visualise several cell layers in one region of interest of the epithelium (Rangrez, Bussau, Ifrit, & Delaney, 2021) (Figure 1.7.). A second generation of handheld CLEs called FIVE2 were introduced in 2021 with improved Z-depth control and better resolution. This CLE technology was used in Optiscan's InVivage^{TM} dental imaging device and Carl Zeiss Meditec's CONVIVO^{TM} for use in neurosurgery (Rangrez, Bussau, Ifrit, & Delaney, 2021).

CLE is but one of the several optical imaging techniques used for medical diagnosis. Optical coherence tomography (OCT) is another similar technology that has been used to produce images of head and neck tissue by measuring the intensity of backscattered light within the tissue (D. Huang et al., 1991). For handheld use within the human body flexible OCT probes provide resolutions as small as 10 μm and penetration depths of 2 mm into soft tissue. (Kraft et al., 2008). OCT technology was aimed at recognising the integrity of the basement membrane of the epithelium and identifying early signs of invasive disorders such as cancer by being sensitive to the change in contrast between the epithelium and underlying connective tissue (Kraft et al., 2008). Another similar imaging modality is photo-acoustic imaging (PAI). PAI induces short laser pulses in the tissue to create thermoelastic expansions within tissue structures that have been targeted by these pulses. These pulses can be detected and imaged up to a depth of 5cm with a resolution of 50 μm (Beard, 2011). While PAI has a greater imaging depth compared to OCT its resolution is higher, indicating it is not as proficient at imaging smaller structures.

This pattern of greater imaging depth coming at the cost of higher imaging resolution is seen across all medical imaging. As seen in Figure 1.8. adapted from Volgger et al. (2013) conventional positron emission tomography (PET) and computed tomography (CT)/magnetic resonance imaging (MRI) scans provide increased imaging depth within tissue at the cost of being unable to provide microscopic resolution (Volgger, Conderman, & Betz, 2013b). This is where the CLE technology shines as it can provide exceptionally small resolution images were characteristic landmarks in individual epithelial cell contents, nuclei, cell borders and smaller extracellular components can be clearly demarcated (Figure 1.8.).

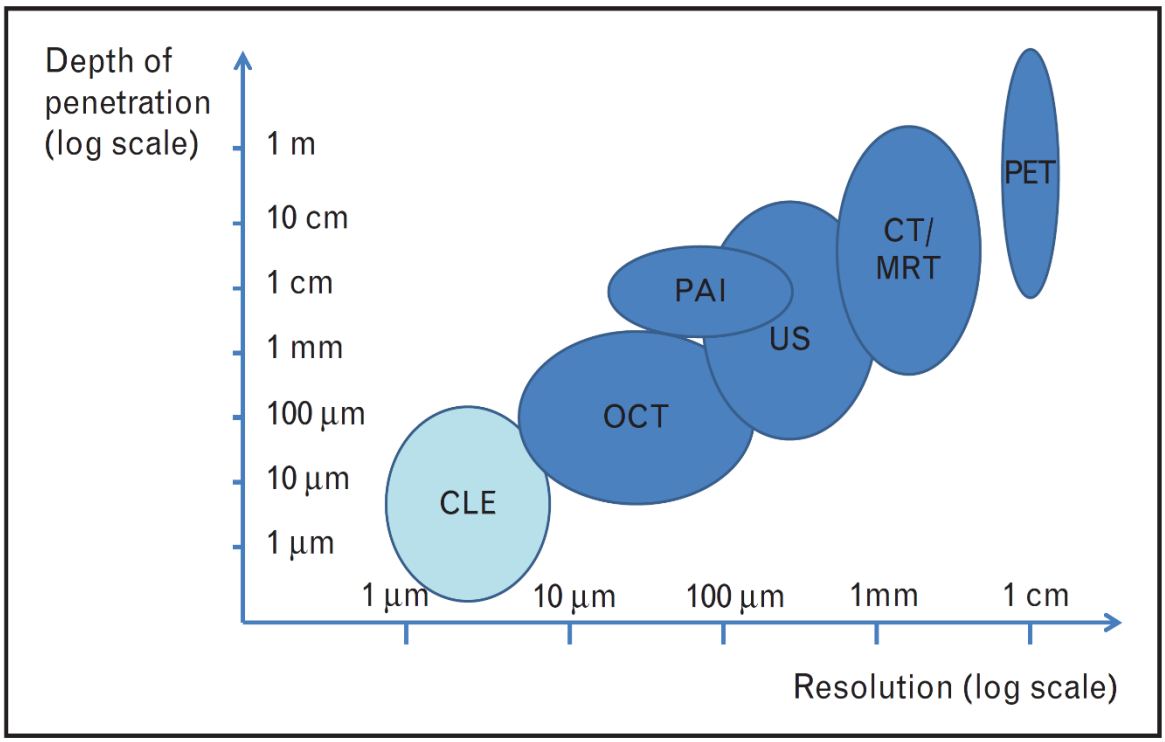


Figure 1.8. Comparison of imaging depth (y-axis) to resolution (x-axis) of various medical imaging modalities. Adapted from Volgger et al. (2013)

A few commercial CLE systems are clinically certified by approval from the Communaute´ Europe´enne and Food and Drug Administration (FDA) as this technology is being adapted for use by clinicians. Two different CLE system have been used in clinical endoscopy, the probe-based CLE and the endoscope-based CLE. The ISC 100 Endomicroscope (Pentax Life Care, Tokyo, Japan) are endoscope-based where the CLE system is integrated into the distal tip of a conventional endoscope. CLEs such as the Pentax system provide imaging depth of up to 250 µm with lateral resolution as small as 0.7 µm (Joey M Jabbour, Meagan A Saldua, Joel N Bixler, & Kristen C Maitland, 2012). The probe based systems on the other hand such as OptiScan Imaging (Mulgrave, Australia), Cellvizio (Mauna Kea Technologies, Paris, France) and VivaScope

(Munich, Germany) provide lateral resolution in the range of 0.55 to 1.25 μm with an axial resolution of close to 5.1 μm (LeCun, Bengio, & Hinton, 2015a; Rangrez, Bussau, Ifrit, & Delaney, 2021; Yap et al., 2023).

These miniaturised CLE devices have been applied for diagnostic purposes in the gastrointestinal (GI) tract, urinary tract, cervical tissue, and respiratory system (Joey M Jabbour et al., 2012). The most widely reported application of CLEs is in the GI tract for real-time microscopic visualization and detection of conditions such as Barrett's esophagitis, inflammatory diseases, and early-stage malignancies (Pilonis, Januszewicz, & di Pietro, 2022). A review by Volgger et al. (2013) opened discussion regarding the potential of confocal microscopy in head & neck disorders. This review noted the lack of studies on CLE use in head and neck cancer compared to other fields despite the proven proficiency of CLEs in identifying dysplastic and microinvasive lesions (Volgger et al., 2013b). This was followed by a review by Lucchese et al. (2016) that recognised the potential of reflectance confocal imaging in intra oral diagnosis (Lucchese et al., 2016b). Maher et al. (2016) summarised the in vivo applications of confocal microscopes for oral mucosal pathologies. They recognised that the quality of evidence in the studies analysed was poor and the data was mainly limited to small descriptive studies (N. Maher et al., 2016). The use of confocal microscopy in the diagnosis of oral cancer was reviewed by Ramani et al. (2022) as described in Chapter 2 of this dissertation (Ramani et al., 2022).

1.5. Fluorescence confocal microscopy

Building on the principles of confocal microscopy, which provides high-resolution optical sections of tissue by eliminating out-of-focus light, fluorescence confocal microscopy (FCM) enhances image contrast and specificity through the use of fluorescent dyes (Ragazzi et al., 2014). By labelling cellular structures with fluorophores, FCM produces detailed grayscale images that highlight areas of high cellular density. This added contrast makes FCM particularly valuable in pathology, where accurate identification of neoplastic versus normal or reactive tissue is critical (Ragazzi et al., 2014). Two classic dyes, acriflavine and fluorescein, have long been used as topical contrast agents to visualize cells and tissues under fluorescent illumination. Acriflavine, an acridine-derived dye, preferentially stains nuclei with a yellow-green fluorescence, whereas fluorescein, a xanthene dye, diffuses through tissues and emits bright green fluorescence (Piorecka, Kurjata, & Stanczyk, 2022; Robertson, Bunel, & Roberts, 2013). Both compounds have rich histories in science and medicine and remain relevant in modern imaging techniques.

1.5.1. Acriflavine

Acriflavine is an acridine dye derived from coal tar used in the form of an orange-brown, odorless powder that readily stains organic materials (Piorecka et al., 2022). It was first synthesised in 1912 by German scientist Paul Ehrlich as a topical antiseptic, it was extensively used during World War I to treat

Figure 1.9. A bottle of acriflavine from 1952 used by Australian nurses in World War 2 (image borrowed from Victorian Collections: https://victoriancollections.net.au)

wounds and combat the parasites causing African trypanosomiasis (sleeping sickness) (Piorecka et al., 2022).

Its ready availability and broad antimicrobial action made it valuable before modern antibiotics. For example, Australian field nurses in WWII still carried acriflavine in their kits (Figure 1.9.). Acriflavine is a cationic (positively charged) dye that intercalates into nucleic acids, giving it an affinity for Deoxyribonucleic acid (DNA)/Ribonucleic acid (RNA). This dye is water-soluble and can penetrate cells due to its small size and amphipathic structure (Tubbs, Ditmars Jr, & Van Winkle, 1964). While it's medical uses are largely of historical interest, its fluorescence properties have kept it relevant in biomedical research.

Chemically, acriflavine is a combination of two compounds: 3,6-diamino-10-methyl-acridine chloride (trypaflavine) and 3,6-diaminoacridine (proflavine) (Piorecka et al., 2022) (Figure 1.10.).

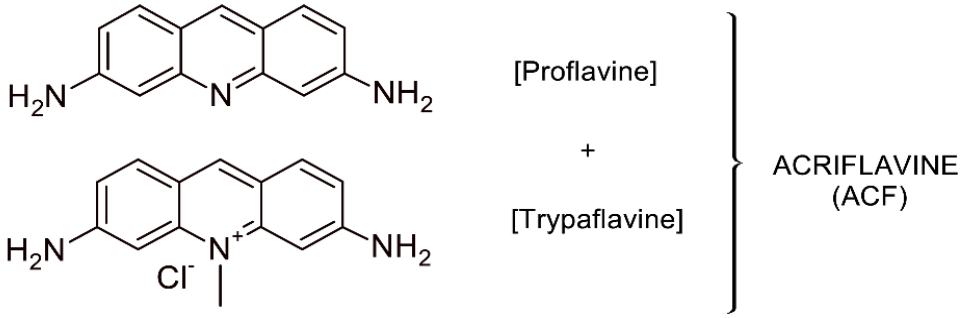


Figure 1.10. Chemical structure of acriflavine as adapted from Pioreck et al. (2022)

Importantly for microscopy, acriflavine's DNA intercalation yields bright nuclear fluorescence, providing high contrast between cell nuclei and the surrounding cytoplasm (Tubbs et al., 1964). It has an excitation peak ~460 nm (blue) and emission around 515 nm (green) making it convenient for fluorescence imaging with standard blue-light sources (Prieto, Powless, Boice, Sharma, & Muldoon, 2015). Unlike targeted immunofluorescent dyes, acriflavine provides non-specific "pan-cellular" staining that highlights general cell architecture (Prieto et al., 2015). This simplicity makes it attractive for point-of-care microscopy and low-resource settings.

1.5.2. Fluorescein

Fluorescein is an organic dye of the xanthene family (a type of triarylmethane dye) (Robertson et al., 2013). In pure form it is a dark orange-red crystalline powder, but it's famed for the brilliant green fluorescence it emits in solution (Figure 1.11.).

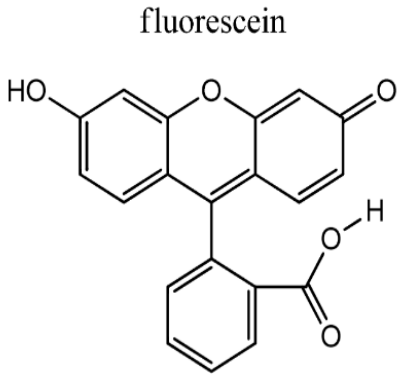


Figure 1.11. Chemical structure of fluorescein as adapted from Robertson et al. (2013)

Fluorescein was first synthesized in 1871 by Adolf von Baeyer (Ragazzi et al., 2014). The dye's optical properties depend on pH. In neutral or basic conditions, its anionic form fluoresces green, whereas in acidic form it may appear more yellow-orange (Figure 1.12.).

Figure 1.12. A representation of fluorescein dye in a beaker of water (image borrowed from the Tintex website: https://tintex.com.au/product/drain-dye-fluorescein/)

In microscopy, fluorescein's value lies in its role as a general contrast agent that can be applied topically or systemically to visualize structures under blue light. Fluorescein has an excitation peak around 494 nm (blue light) and an emission

peak near 521 nm (green light) (Robertson et al., 2013). In fluorescence microscopy of tissues, fluorescein is often used as a broad-spectrum pan cytoarchitectural stain. The fluorescein eye stain test has been a standard diagnostic method for clinicians to quickly visualise abrasions, foreign bodies, or infections on the transparent cornea, which would be hard to see otherwise (Mocan & Irkec, 2007).

Unlike acriflavine, fluorescein does not bind strongly to any one cellular component; instead, it tends to distribute in the extracellular space or bind weakly to plasma proteins. This makes it a versatile tracer for highlighting anatomical structures and fluid flow, albeit with less intrinsic specificity for organelles.

1.6. Microscopy image analysis

Microscopy images of oral mucosa (e.g., histological sections of oral tissues) have been quantitatively analysed using traditional image processing techniques (Banerjee, Kamath, Lavanya, Shruthi, & Deepa, 2015).

Traditional image analysis pipelines typically begin with preprocessing to enhance important features and reduce noise. For example, global or adaptive thresholding can separate tissue from background and even serve as a simple noise filter (Sieracki, Reichenbach, & Webb, 1989). Basic filters (mean, median) or morphological operations are often applied to smooth the image and remove artifacts, which in turn improves subsequent thresholding (Devi & Patil, 2020).

A crucial step is distinguishing cells (especially nuclei) from surrounding tissue. In oral epithelium images, segmentation can be done by pixel-level classification (identifying nuclear pixels vs. background) or by explicit boundary detection Global thresholding (e.g., Otsu's method, watershed) (Kulwa et al., 2019) (Figure 1.13.). For instance, nuclei which appear dark blue/purple in haematoxylin & eosin (H&E) staining, can be isolated by thresholding the blue hematoxylin channel. However, a single threshold may miss faintly stained nuclei or include noise.

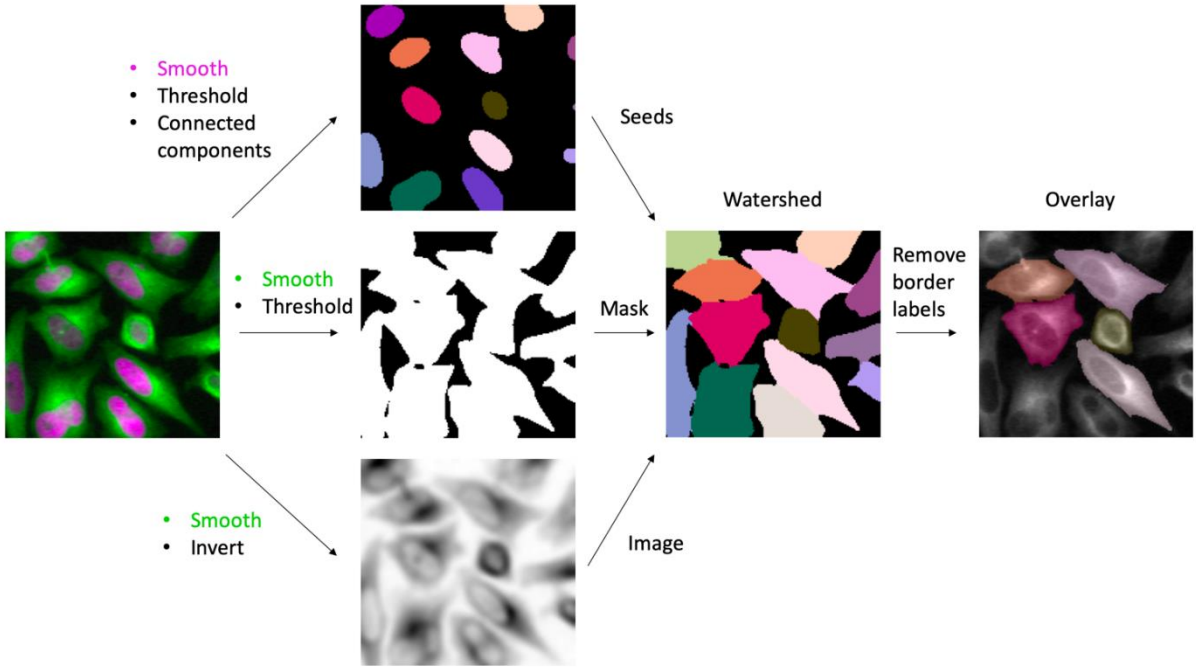


Figure 1.13. Traditional microscopy image analysis workflow for segmenting cells (image borrowed from neubias.github.io)

After segmentation, the next step is featuring extraction from identified objects. Traditional morphometric features quantify cell/nucleus size, shape, and arrangement. Common morphological features include area, perimeter, circularity/compactness, major-minor axis lengths, aspect ratio, symmetry, and concavity of nuclei. These features capture differences in cell morphology. Textural features (often computed via gray-level co-occurrence or other matrix methods) measure intensity variation (smoothness, contrast, etc.) within the tissue or nucleus (Gamarra, Zurek, San-Juan, Eng, & Norte, 2017).

In oral pathology research, such features have been used to distinguish normal mucosa from dysplastic or cancerous tissue. Banerjee et al. (2015) reviewed both object-based features (morphology, structure) and pixel-based texture measures that have been incorporated into diagnostic algorithms for oral histopathology (Banerjee et al., 2015). Using these quantitative features, early computer-aided diagnosis studies could classify tissue samples by disease state using statistical or machine-learning classifiers. Several open-source software tools have become popular for implementing the above image analysis techniques in medical and dental microscopy. Notably, ImageJ/Fiji, CellProfiler, and QuPath are widely used in the last decade for analysing histopathology and immunohistochemistry (IHC) images of tissues (Belaldavar, Angadi, & Mudenagudi, 2024; Schindelin, Rueden, Hiner, & Eliceiri, 2015; Stirling et al., 2021).

Over the last decade, traditional image processing has proven to be a powerful asset in the analysis of oral mucosa microscopy images. AI extends traditional image analysis from rule-based pipelines to data-driven, adaptable systems that can generalise across patients, labs, and staining protocols. In oral histopathology and immunohistochemistry, AI is helping shift from qualitative visual grading to quantitative, reproducible diagnostics, but it builds on a foundation laid by decades of classical image processing.

1.7. Artificial intelligence and computer vision

The advent of ‘intelligent’ machines, once described from the imagination of science fiction visionaries is now closer to reality with the growth of artificial intelligence (AI). AI serves as the overarching domain dedicated to creating systems that emulate human cognitive functions such as reasoning, learning, perception, and problem-solving (Shneiderman, 2020).

Within this broad field, machine learning (ML) emerged as a subset that focuses on developing algorithms that enable systems to learn from data and improve performance over time without explicit programming. Unlike traditional AI approaches that relied on pre-defined rules, ML uses statistical methods and models to identify patterns and make predictions. Examples include supervised learning, unsupervised learning, and reinforcement learning that have been applied to applications such as predictive analytics and recommendation systems (Bishop & Nasrabadi, 2006).

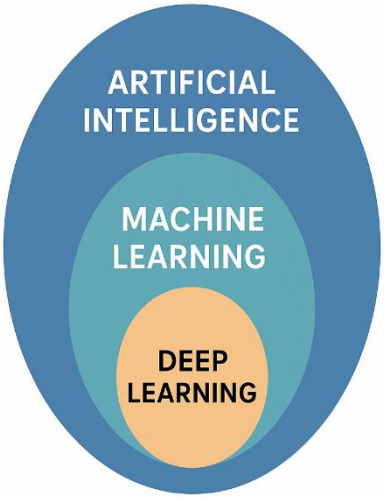


Figure 1.14. Deep learning is a specialisation of machine learning which is a subset of artificial intelligence

A further specialisation within ML is deep learning (DL) that employs artificial neural networks to analyse data hierarchically through multiple processing layers, mimicking the human brain's structure and function (Hinton, Osindero, & Teh, 2006) (Figure 1.14.). DL distinguishes itself from other ML approaches by automating feature extraction, allowing neural networks to learn high-level abstractions from raw data. For instance, in image recognition, initial layers detect basic features like edges, while deeper layers recognize complex patterns such as objects or faces.

The development of DL techniques has been bolstered by the availability of large datasets and advanced computational power, leading to significant breakthroughs in fields such as natural language processing, autonomous vehicles, and medical imaging (LeCun et al., 2015a). Together, these fields form a nested hierarchy: AI encompasses all "intelligent" systems, ML represents data-driven learning within AI, and DL resides as a subset of ML that leverages multi-layered neural network models (Figure 1.14.). The relationship represents the progression from general intelligence systems to highly specialised, data-intensive methodologies that power many state-of-the-art applications in technology.

There are 2 main types of ML models that are based on the type of problem: supervised and unsupervised learning. In supervised ML, models are trained using labelled datasets where input data samples are paired with an annotated 'ground truth' labels that represents the correct target output. The goal of supervised learning is to train models to learn how to map specific inputs to targeted outputs. The applications of supervised learning range from regression problems where numbers need to be predicted, such as predicting house prices, to classification problems where a label needs to be assigned to input data, such as spam email detection (LeCun et al., 2015a). Some popular supervised ML models are linear regression, logistic regression, support vector machines (SVM), decision trees, random forests, and convolutional neural networks (CNN). This supervised learning approach has been found to produce highly accurate models that generalise well on labelled training and test data (Jiang, Gradus, & Rosellini, 2020).

Unsupervised learning on the other hand deals with unlabelled data. The goal of this approach is to identify underlying patterns or structures in data by grouping similar data points or reducing dimensionality. There is no target output variable provided to the models while training. Some applications of this concept are data exploration in marketing to understand customer behaviour, anomaly detection by banks for detecting fraud, and recommendation systems by e-commerce websites to improve consumer interaction (James, Witten, Hastie, Tibshirani, & Taylor, 2023).

Deep learning has the ability to learn complex tasks by transforming input data at every step of its learning process into simpler and more abstract forms. DL in the context of understanding images functions at the pixel level. The initial layers of these algorithms assess the input images as arrays of pixels and look for the most basic features that can be identified, such as the presence and location of edges of objects in the image. Edges are usually defined by the sharp change in pixel brightness, colour or contrast between adjacent pixels. The succeeding layers attempt to spot particular arrangements of these edges.

Eventually the later layers in the DL model identify these motifs from the arrangement of edges and draw parallels with parts of familiar objects.

The ‘familiarity’ these DL models have with objects in the real world is based on context provided by human DL practitioners that develop these models. While humans design the foundation of how these models learn, the actual content and context learned by these models are not determined by humans, which makes this technology flexible and adaptable to a range of tasks. DL models created thus have proved to be exceptionally capable at identifying intricate patterns in high dimensional data. It has demonstrated its excellence in several domains of science, business, and government (LeCun et al., 2015a).

Computer vision is a field of AI that enables computers to extract meaningful information from digital images, videos and other visual data followed by taking actions or making recommendations. This field of study focuses on providing machines the ability to see and understand the world around us. In theory computer vision aims to mimic human vision by understanding context from visual information by assessing a large number of examples. Humans have a head start in such matters due to our lifetime of experience in telling objects apart, gauging how close or far they are, assessing their size, and making sense of the objects we visualise and even pointing out when something appears to be wrong in images. The primary tasks carried out by computer vision models are image classification, object detection, and image segmentation. These tasks focus on different properties of the target image to be analysed:

1. **Image Classification**

Image classification involves assigning a predefined label or category to an entire image through making predictions. This task enables systems to accurately identify whether an image contains a specific object, such as a dog, an apple, or a person's face. It serves as the basis for many applications, such as facial recognition, customer behaviour recognition, autonomous vehicles as well as identifying diseases on medical images. The use of CNNs has significantly improved the accuracy of image classification tasks, particularly on large-scale datasets like ImageNet (Deng, Dong, Socher, Li, Kai, et al., 2009).

2. **Object Detection**

Object detection extends image classification by identifying and localising specific objects within an image. This involves not only determining whether a particular class of object is present but also pinpointing its exact location within the image or video frame. Object detection is critical for numerous industrial and commercial applications, including quality control in manufacturing where

systems can identify defective items, or safety systems in vehicles that detect obstacles on the road (He, Zhang, Ren, & Sun, 2016).

3. Image Segmentation

Image segmentation takes object detection a step further by partitioning an image into segments, where each pixel is assigned, a label corresponding to the object or region it belongs to. This pixel-level classification provides a detailed understanding of the image, enabling more precise analysis. Segmentation can be further divided into two categories:

i) Semantic segmentation that classifies all pixels of an image into categories
ii) Instance segmentation that identifies individual instances of objects within the same class

Applications of segmentation include medical imaging, where it is used to delineate anatomical structures with high precision, and autonomous vehicles, that rely on segmentation to distinguish between different road elements, such as lanes, pedestrians, and traffic signs. Notable frameworks for image segmentation include U-Net, widely used in medical image analysis, and Mask R-CNN that extends Faster R-CNN to perform instance segmentation (Ronneberger, Fischer, & Brox, 2015).

By integrating these three tasks, classification, detection, and segmentation, modern computer vision systems achieve a comprehensive understanding of visual data for a wide range of applications.

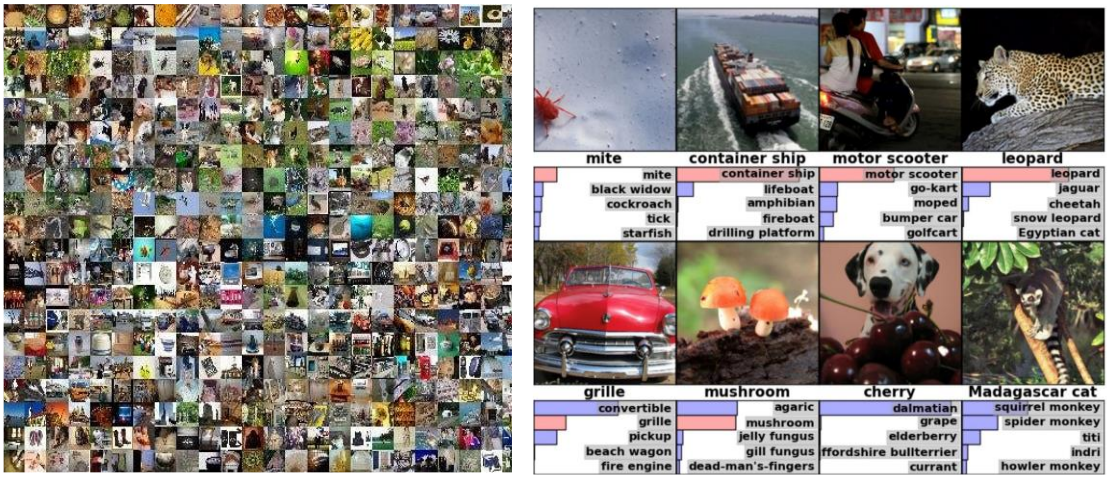
The evolution of machines to interpret visual data spans over six decades, beginning in 1959 when Hubel and Wiesel conducted landmark experiments demonstrating how the brain responds to visual stimuli. Their findings showed that image processing starts with simple shapes such as edges and lines (Hubel & Wiesel, 1965). This discovery laid the groundwork for image processing and computational models of vision. Around the same time, the development of digital image scanning technologies enabled computers to digitize visual data, with early advancements focused on converting 2D images into 3D forms.

The introduction of Optical Character Recognition (OCR) technology in the 1970s was another significant milestone with companies such as Kurzweil Computer Products releasing the first OCR system capable of recognizing multiple fonts (Goodrich, Bennett, De L'aune, Lauer, & Mowinski, 1979). Building on this, Intelligent Character Recognition (ICR) emerged to process handwritten text using neural networks, further expanding applications in document automation and handwriting recognition (LeCun, Bottou, Bengio, & Haffner, 1998).

In the 1980s, neuroscientist David Marr revolutionized understanding of visual processing through his hierarchical model of vision. His seminal work, "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information" described algorithms for detecting edges, corners, and curves (Marr, 2010). Concurrently, Kunihiko Fukushima developed the Neocognitron, a neural network featuring convolutional layers, a precursor to modern deep learning architectures (Fukushima, 1980).

This era also saw the establishment of datasets such as ImageNet, released in 2010, that provided a standardized, annotated database essential for training deep learning models (Deng, Dong, Socher, Li, Kai, et al., 2009). The breakthrough moment arrived in 2012 when AlexNet, a deep learning model developed by Geoffrey Hinton, Alex Krizhevsky, and Ilya Sutskever, reduced image recognition error rates dramatically, leading the way for modern computer vision applications (Krizhevsky, Sutskever, & Hinton, 2012). This model was developed on the ImageNet database that was one of the first large-scale, accurate, and diverse repositories of hand curated images of objects around the world and has been a critical resource for the development of advanced, large scale image understanding algorithms (Deng, Dong, Socher, Li, Kai, et al., 2009). This database, originally conceived by the pioneer Prof. Fei-Fei Li in 2006, is now a freely accessible repository containing over 14 million labelled images, organized into more than 20,000 categories of objects, such as 'beach ball' and 'strawberry' (Figure 1.15.).

Since its inception, the ImageNet database has been used for an annual open software competition called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) that aims to highlight the very best image recognition algorithms and models trained and benchmarked on the ImageNet database via supervised ML principles (Figure 1.15.) (Deng, Dong, Socher, Li, Kai, et al., 2009). This has led to the development of large number of models based on this database. The top performing models in the ImageNet challenge over the years were CNN models, such as AlexNet (Krizhevsky et al., 2012).



A B

Figure 1.15. The ImageNet database for development of large-scale accurate image identification models. A) A sample panel of the 14 million images; B) Examples of image labelling provided to help develop the models (adapted from Deng et al., 2009)

These state-of-the-art CNNs are DL models designed to process data with a grid-like structure, such as images. Inspired by the organization of the visual cortex in the human brain, CNNs are highly effective for visual tasks, leveraging their ability to automatically learn spatial hierarchies of features from raw data. The primary building blocks of a CNN include convolution layers, pooling layers, activation functions, and fully connected layers (Figure 1.16.).

The convolution operation involves applying filters (kernels) to assess unique sections of the image to extract features such as edges, textures, and patterns, generating feature maps. Pooling layers, such as max pooling and average pooling, reduce the spatial dimensions of these feature maps, preserving essential information while lowering computational requirements. Non-linear activation functions like ReLU (Rectified Linear Unit) enable the network to model complex relationships. Fully connected layers, found at the end of the network, combine the learned high-level features to make predictions (Z. J. Wang et al., 2020). A representation of these different types of layers found in the CNN are depicted in Figure 1.16., which has visuals from Wang et al.'s CNN explainer project (Z. J. Wang et al., 2020).

During training, the machine is presented with an image and generates a vector of scores, where each score corresponds to a category. Ideally, the score for the correct category should be the highest, but this is usually not the case initially. To address this, an objective function is calculated to measure the discrepancy (or error) between the predicted scores and the target scores. The machine then updates its internal parameters, known as weights, to minimise this error. These weights are adjustable numerical values that determine how the machine maps inputs to outputs. In a typical deep learning model, there can be hundreds of

millions of these weights that are fine-tuned using vast datasets of labelled examples. To update the weights effectively, the learning algorithm calculates a gradient vector. This vector indicates how the error would change if each weight were slightly adjusted. The weights are then updated in the opposite direction of the gradient to reduce the error in a process called backpropagation. This process allows the machine to iteratively refine its predictions and improve its performance (LeCun et al., 2015a).

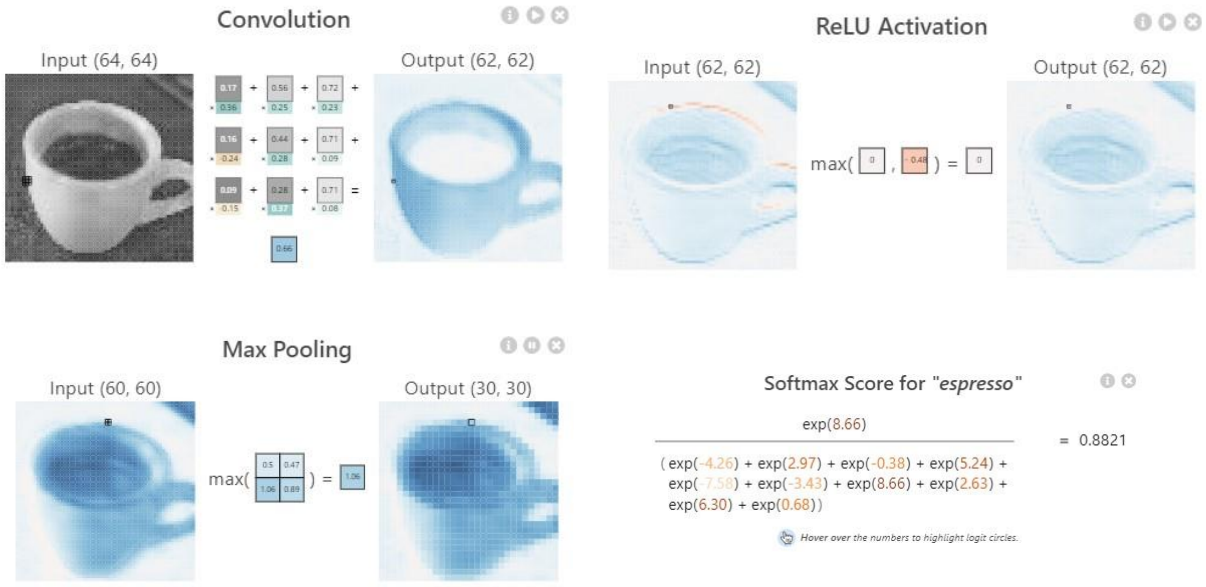


Figure 1.16. Representation of the Convolution, ReLU activation, Max Pooling and SoftMax layers of CNNs using the visualisation from the CNN Explainer project by (Z. J. Wang et al., 2020)

The most widely used algorithm to optimise these weights during model training is the stochastic gradient descent (SGD) algorithm. SGD works by updating model parameters in the direction of the negative gradient of the loss function with respect to each parameter. This process occurs iteratively over small, randomly selected subsets of the training data (mini-batches) so as to reduce computational cost compared to calculating gradients over the entire dataset. While effective, standard SGD can be sensitive to hyperparameter tuning parameters such as the learning rate and may therefore converge slowly or result in the algorithm becoming stuck in local minima (LeCun et al., 1998).

Hyperparameters are user-defined variables that control the training process of a machine learning model. Among the most important hyperparameters in deep learning are epochs and learning rate that significantly influence model performance and convergence. Epochs refer to the number of times the entire training dataset is passed through the model during training. Training a model

for more epochs allows it to learn more patterns from the data, but too many epochs can lead to overfitting, where the model performs well on the training data but poorly on unseen data. Conversely, too few epochs might result in underfitting, where the model fails to capture sufficient patterns in the data. Selecting an appropriate number of epochs often involves monitoring metrics such as validation loss or accuracy and using early stopping techniques to terminate training when performance no longer improves.

Learning rate controls the size of the steps taken during the optimization process to update the model's parameters. It determines how quickly or slowly the model converges toward the optimal solution. A high learning rate can speed up training but may cause the model to overshoot the optimal solution or fail to converge. A low learning rate ensures more precise convergence but may significantly increase training time or result in the algorithm becoming stuck in suboptimal solutions. Together, these hyperparameters require careful tuning to achieve a balance between training time, convergence, and generalization performance. Techniques such as grid search, random search, or the more sophisticated method Bayesian optimization, are often used to find optimal values for these hyperparameters for any given task (Probst, Wright, & Boulesteix, 2019).

One of the top performing CNN models developed during the ILSVRC was the Inception_V3 model. Inecption-V3 is a pre-trained network developed by researchers at Google in 2015 as an evolution of their previous CNN architectures and led to significant improvements in both performance and computational efficiency. One of the main ideas in Inception_V3 is breaking down big tasks into smaller, easier ones. For example, instead of assessing a large section of an image all at once, the model assesses smaller sections one at a time making the process faster but still accurate. Additionally, asymmetric convolutions are employed, where operations are split into two steps, such as a 1x3 convolution followed by a 3x1 convolution. These further decreases computational cost while effectively capturing spatial patterns in the data.

A further optimisation involved efficient grid size reduction, where techniques such as pooling and stride convolutions are used to down-sample feature maps. These methods preserve critical information while reducing the resolution, enabling the model to focus on salient features with minimal information loss. Inception_V3 includes auxiliary classifiers, that serve as additional layers to help improve the flow of gradients during training and reduce the risk of overfitting. The architecture makes extensive use of batch normalisation layers, speeding up training and enhancing the model's ability to generalise to new data.

Additionally, it employs a technique called label smoothing assigning a small probability to incorrect labels during training. This reduces overfitting and

improves the reliability of the model's predictions. Comprising 48 layers, Inception_V3 has a modular design combining different types of convolutions and pooling in parallel to capture a large range of features. This design enables the model to achieve high performance on datasets such as ImageNet with a top-5 accuracy of 93.9% while being computationally more efficient than other models with similar accuracy (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016).

Complex CNN models such as Inception_V3 are made up of millions of parameters that are updated and optimised by learning algorithms and hyperparameters. Developing models of such magnitude and complexity would generally involve a considerable amount of expertise in computer science along with large scale computational resources. In order to prevent these barriers from slowing down innovation in the field of DL, several groups of scientists all over the world are working towards the democratisation of deep learning for everyone. This involves open access to complex models such as Inception_V3, larger scale databases such as ImageNet, and frameworks for fair and safe use of DL technology (Lyu, Li, Nandakumar, Yu, & Ma, 2020).

Modern DL frameworks used to build these models such as PyTorch and TensorFlow have made significant contributions to democratising DL by providing accessible, flexible, and powerful tools to researchers, developers, and AI enthusiasts. PyTorch, developed by Facebook's AI research lab, is a Python-based deep learning framework known for its flexibility, dynamic computation graph, and ease of use. One of its key features is its "eager execution" mode that allows users to debug and modify code on the fly. Its automatic differentiation capabilities ‘AutoGrad’ enable it to handle all the complex differential mathematics under the hood while providing the user the amount of visibility they need (Paszke et al., 2017b). This makes it particularly popular among researchers, as it facilitates rapid prototyping and experimentation. PyTorch also provides a comprehensive ecosystem, including ‘torchvision’ for importing freely available open source CNN models, carrying out image processing tasks, and developing training workflows (Paszke et al., 2019).

The community-driven nature of PyTorch, along with its straightforward application programming interface (API), has made it a preferred choice in academia, accelerating the pace of research in deep learning. PyTorch has contributed to the democratization of DL by lowering the entry barrier for learners, enabling fast experimentation, and offering extensive documentation and tutorials. Furthermore, its adoption in cutting-edge research ensures that state-of-the-art techniques are accessible to the broader community. A prominent movement in the microscopy community towards the

democratisation of DL is Zero Cost Deep Learning for Microscopy (ZeroCostDL4Mic) (von Chamier et al., 2021). ZeroCostDL4Mic is an open-source platform designed to make deep learning more accessible to researchers in microscopy, particularly those without extensive coding skills or access to expensive computational resources. It is implemented using Google Colaboratory, a cloud-based platform that provides free access to graphics processing units (GPU) and tensor processing units (TPU), eliminating the need for specialized hardware. The platform offers a collection of pre-implemented deep learning models tailored for microscopy tasks such as image restoration, segmentation, classification, and object detection. By running entirely on the free tier of Google Colaboratory, it removes financial barriers while supporting a wide range of microscopy applications, such as cell segmentation, noise reduction in live-cell imaging, subcellular structure analysis, and phenotypic classification in drug screening (von Chamier et al., 2021).

Among the DL models offered on the ZeroCostDL4Mic platform is StarDist 2D (Weigert, Schmidt, Haase, Sugawara, & Myers, 2020). StarDist 2D is a deep learning-based framework designed for instance segmentation of objects with star-convex shapes, such as cells or nuclei, in 2D microscopy images. Unlike traditional segmentation methods that predict pixel-wise labels, StarDist uses a geometric approach to segment objects by representing their shapes as star-convex polygons. Each object is defined by radial distances from a central point to its boundary, allowing for efficient segmentation of overlapping and irregularly shaped objects. StarDist predicts both the radial distances and object probabilities for each pixel, enabling the model to simultaneously identify the objects centre and reconstruct their shapes. This approach makes it particularly effective for microscopy data, where objects are often densely packed and exhibit diverse morphologies. StarDist 2D provides pretrained models that can be directly applied to typical fluorescence microscopy datasets, while also allowing users to fine-tune these models for improved performance on custom data (Weigert et al., 2020).

Cellpose is another deep learning-based segmentation algorithm designed specifically for microscopy images, developed by Stringer et al. (2021) at the Howard Hughes Medical Institute (Stringer, Wang, Michaelos, & Pachitariu, 2021). It's widely recognised for its ability to accurately segment cells and nuclei across diverse imaging modalities, cell types, and staining protocols, without requiring retraining or large annotated datasets. Cellpose generates a probability map of where cells are and flow field vectors which indicate how to move from any point in the cell to the centre. The segmentation is then computed by simulating particle movement along these flow fields until pixels “converge” into object centres. This allows high precision in boundary detection and better

separation of touching cells compared to traditional algorithms such as watershed (Pachitariu & Stringer, 2022).

Methods such as StarDist and Cellpose are building upon the foundations laid down by traditional microscopy image analysis methods by harnessing the power of artificial intelligence architectures and modern high performance computing hardware to push the boundaries of image analysis.

1.8. Deep learning in medical diagnosis

The advent of AI in medical imaging has been a transformative leap in various medical fields such as dermatology, ophthalmology, and radiology, with accuracy values similar or better than that of experienced clinicians (Sun et al., 2020).

AI technologies, particularly machine learning and deep learning, have significantly enhanced the ability to analyse and interpret complex medical images, providing precise and timely diagnostics for diseases ranging from cancer to neurological disorders (Phillips, 2020). ML methods have entered the field of precision medicine, which refers to the approach of understanding diseases at the individual patient level in order to provide bespoke treatment plans for each patient based on advanced technologies (Phillips, 2020). Machine learning has been making its way into precision medicine due to its ability to rapidly utilise complex data from multiple sources for predicting disease outcomes for both individuals and populations (MacEachern & Forkert, 2021).

Dentistry is a field that has wide range of clinical tasks that lend themselves well to being augmented by ML analysis. AI applications are rapidly entering dentistry with the promise of making diagnosis and treatments faster, and personalised. However, their robustness, generalisability, transparency and reliability are questionable to due poor reporting (Schwendicke et al., 2021). A recent review of ML research in dentistry by Arsiwala-Scheppach et al. (2023) revealed a wide variety of ML tasks and data types being studied in dentistry including image data from radiographs, scans, and photographs (Arsiwala-Scheppach, Chaurasia, Mueller, Krois, & Schwendicke, 2023). This variability, coupled with data primarily sourced from single centres representing limited populations, adversely impacts the generalizability of the results. Outcome measures varied widely, with some studies not clearly specifying the level (e.g., patient-level, tooth-level, or surface-level) at which outcomes were assessed, hindering cross-study comparisons (Arsiwala-Scheppach et al., 2023).

Strategies for defining ground truth (reference tests) varied widely, and many studies provided inadequate details about their methodologies. In cases where only one human annotator was used as the reference, results are susceptible to variability and bias (Arsiwala-Scheppach et al., 2023). Many studies displayed deficiencies in conducting and reporting, including issues with biases, data leakage, and overfitting. Validation of results on external datasets was often missing, limiting generalizability. There was a lack of exploration into why models failed to generalise and how to improve them. Most studies focused on

demonstrating that ML models could learn and predict but paid insufficient attention to understanding the underlying mechanisms, clinical needs, and safeguards required for ML in dentistry (Arsiwala-Scheppach et al., 2023).

The World Health Organization (WHO) - International Telecommunication Union (ITU) – World Intellectual Property Organization (WIPO) Global initiative on AI for Health, Topic group – Oral Health has been working on guidelines for the safe and ethical use of AI to address these concerns. Researchers from this group published their Delphi process derived checklist for researchers and readers of AI research in dentistry to promote AI literacy among dentists and researchers (Schwendicke et al., 2021). A paper on core outcome measures for computer vision studies in dentistry also identified this heterogeneity in reporting performance metrics across the current literature (Büttner, Rokhshad, et al., 2024).

A recent review of studies employing AI techniques for detecting oral mucosal lesions from clinical photographs noted that despite the heterogeneous approach to dental AI methodology and reporting majority of the studies reported high classification performance (Rokhshad et al., 2024). The pooled sensitivity and specificity of detecting cancerous lesions across 15 studies reviewed was 0.90 (95% confidence interval 0.85-0.94) and 0.92 (95% confidence interval 0.89-0.95), respectively (Rokhshad et al., 2024). Studies evaluating the performance of AI in classifying, detecting, or segmenting oral lesions from clinical photographs reported varied outcomes. Overall accuracy ranged from 74% to 100%, sensitivity from 0.58 to 0.99, specificity from 0.40 to 1, and AUC from 0.51 to 0.99 (Rokhshad et al., 2024). For image classification, the range for accuracy was 74% to 100%, sensitivity 0.79 to 1, and specificity 0.60 to 0.97 (Rokhshad et al., 2024). Object detection studies showed accuracy of 91% to 95%, sensitivity at 0.95, and specificity between 0.89 and 0.93 (Rokhshad et al., 2024). Segmentation models demonstrated accuracy between 85% and 95%, specificity from 0.85 to 0.87, and sensitivity ranging from 0.57 to 0.96 (Rokhshad et al., 2024).

These findings highlight the variability in AI performance depending on the task and model used and underscore the need for robust methodology and clear reporting for transparency and reproducibility.

1.9. AI for early diagnosis of oral cancer

The morbidity associated with oral cancer can be mitigated by early detection. It is preferable that associated lesions are discovered in the early stages of progression before malignant transformation. This early identification of oral potentially malignant disorders (OPMD) can lead to higher survival rates and better monitoring of the progression (McCullough et al., 2010a). Routine oral examinations carried out by dentists are the most common diagnostic method used.

However, there are some limitations of clinical examination on detection of cancerous or precancerous lesions (Kujan et al., 2007). It is more common for patients to report with advanced stages of oral cancer at the time of diagnosis. At the early stages of oral cancer, the survival rate can be as high as 80%. However, for patients at later stages survival rates can drop to around 20% (Isaäc Van der Waal, 2013). It has been noticed that doctors tend not to associate presented oral symptoms with the possibility of a neoplastic process that could lead to delays in diagnosis (Idrees, Halimi, Gadiraju, Frydrych, & Kujan, 2024). This indicates the prevalence of subjectivity while diagnosing OPMDs and oral cancer.

Rutkowska et al. (2020) found a mean oral cancer diagnostic delay of 222 days in their study exploring delays in oral cancer treatment. The authors noted that some of the reasons for delays by professionals is inadequate clinical examination, supplemented by a low index of suspicion, coupled with a lack of familiarity with the disease. The authors also pointed out that patients tend to downplay the initial symptoms associated with oral cancer with one of the reasons being that these symptoms do not adversely affect their functionality to a large extent (Rutkowska, Hnitecka, Nahajowski, Dominiak, & Gerber, 2020).

Keinanen et al. (2024) in their retrospective study noted that patients actively seeking care for OSCC are 6.6 times more likely to have their diagnosis delayed. The most common challenge clinicians faced was related to tumour identification and the most common treatment attempt made by clinicians before referral was tooth extraction (Keinänen, Uittamo, & Snäll, 2024). In a disease like oral cancer, where survival rates plummet in the final stages of the disease these delays could significantly impact quality of life as well as longevity due to increased morbidity and mortality.

AI may be able to provide rapid and accurate solutions. In oral medicine, AI models can be utilized for various tasks such as disease diagnosis, treatment planning, and outcome prediction. For instance in the diagnosis of oral diseases neural networks can be trained on large datasets of macrographs, micrographs,

histopathology slide images and other sources of patient data (Ramani et al., 2022). These networks can learn complex patterns and features from this data, enabling them to make accurate diagnoses with high sensitivity and specificity (LeCun et al., 2015a). Upon analysing patient data such as medical & dental histories, genetic information, and imaging results, these deep learning models could recommend personalized treatment strategies (Bellando-Randone et al., 2021). They can also predict treatment outcomes, helping clinicians make informed decisions about patient care.

The capabilities of these models in oral pathology diagnosis are becoming increasingly prevalent in the literature with a recent systematic review showing that about 22% of classification machine learning papers in dentistry are related to oral medicine and pathology (Arsiwala-Scheppach et al., 2023). Camalan et al. (2021) validated a method of classifying clinical photographs of oral epithelial dysplasia using an Inception-Resnet-V2 model with an accuracy of up to 90.9% (Camalan et al., 2021). Tanriver et al. (2021) used image segmentation (instance and semantic), object detection and classification experiments using a range of deep learning models such as U-Net, Mask R-CNN and DenseNet-161 to successfully identify benign oral lesions, oral squamous cell carcinoma (OSCC) and oral potentially malignant disorders (OPMD) such as leukoplakia, erythroplakia, and submucous fibrosis (Tanriver, Soluk Tekkesin, & Ergen, 2021).

Shamim et al. (2022) recently developed a ResNet50 deep learning model that demonstrated the capability to differentiate among five varieties of tongue lesions, specifically: hairy tongue, fissured tongue, geographic tongue, strawberry tongue, and oral hairy leukoplakia, with an impressive average classification accuracy of 0.97 (Shamim et al., 2022). Jubair et al. (2022) developed their own version of the EfficientNet-B0 model to detect benign and malignant lesions or oral cancer using clinical images with an accuracy of 85% (Jubair et al., 2022). As AI models are being integrated into clinical diagnosis software packages such as MeMoSA® and MouthMap™, education and training in using such technological advancements along with ongoing professional development would be an important goal for oral medicine and pathology professionals (Welikala et al., 2020; Yap et al., 2023).

Despite AI models being exceptionally powerful and quick, they come with a few drawbacks. Interpretability is a major concern with several large-scale models using modern CNNs. They are often treated as black boxes as the exact mechanisms are often lost within the complexity of these models. Extracting meaningful insights as to how these models reach their decisions is challenging (LeCun et al., 2015a). Another consideration is the ethical implications of creating certain AI models since there are concerns regarding deception,

manipulation or coercion that could adversely affect human autonomy (Laitinen & Sahlgren, 2021). To account for this 11 ethical principles have recently been suggested that need to be kept in mind when designing AI models, including diversity, transparency, wellness, privacy protection, solidarity, equity, prudence, law and governance, sustainable development, accountability, and responsibility, respect of autonomy and decision-making (Rokhshad et al., 2023). Challenges in AI research within oral medicine stem from issues such as small training datasets, unclear data generation processes, and ambiguous data annotation strategies, alongside uncertainties in model selection and validation methods. Research groups around the world including the WHO-ITU-WIPO Global initiative on artificial intelligence for health and the Oral Medicine and Oral Cancer (OMOC) group in Melbourne are utilising robust scientific protocols to surmount these challenges (Schwendicke et al., 2021).

The transformative potential of AI in oral pathology and medicine is profound, promising enhanced diagnostic accuracy, personalised treatment plans, and streamlined patient care. In the future, the ongoing evolution of AI in this field holds the potential for even greater advancements, including novel diagnostic tools, more efficient workflows, and improved patient outcomes through integrated AI-driven solutions.

Despite advances in artificial intelligence for oral lesion detection, most models remain confined to controlled datasets with limited clinical validation and interpretability, leaving uncertainty about their practical impact in real-world dental settings. Additionally, current AI research has yet to deeply validate or integrate novel in vivo confocal microscopy technology, which offers real-time cellular-level imaging with significant diagnostic potential. This gap highlights the need to explore how AI-assisted diagnostic tools can augment accuracy, confidence, and diagnostic speed in identifying oral potentially malignant disorders. It is hypothesised that integrating machine learning systems into oral cancer screening with fluorescence in vivo confocal microscopy would facilitate rapid and accurate detection of oral potentially malignant disorders and oral squamous cell carcinoma.

1.10. Aims and hypotheses

The overall aim of the work in this dissertation was to evaluate the performance of machine learning with fluorescence in vivo captured confocal laser endomicroscopy for the detection of oral potentially malignant disorders and oral squamous cell carcinoma. The overall dissertation research hypothesises that machine learning with fluorescence in vivo captured confocal laser endomicroscopy of the oral mucosa detects oral potentially malignant disorders and oral squamous cell carcinoma with high precision. The background information provided in this chapter point to a project that incorporates the following aims and hypotheses.

1.10.1. Chapter 2 - Systematic review of confocal microscopy in oral cancer diagnosis

Aim: To summarise and evaluate the evidence on the utility and performance of confocal microscopy in the diagnosis of oral potentially malignant disorders and oral squamous cell carcinoma.

Null Hypothesis (H_0): That studies do not utilise confocal microscopy for diagnosing oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That studies utilise confocal microscopy for diagnosing oral potentially malignant disorders and oral squamous cell carcinoma.

1.10.2. Chapter 4 - Quality filtering of confocal micrographs

Aim: To develop a CNN model that can identify diagnostic quality fluorescence in vivo confocal micrographs for downstream diagnostic triage by filtering out poor quality data.

Null Hypothesis (H_0): That a CNN model cannot accurately and rapidly identify fluorescence in vivo confocal micrographs of high diagnostic quality.

Alternative Hypothesis (H_1): That a CNN model can accurately and rapidly identify fluorescence in vivo confocal micrographs of high diagnostic quality.

1.10.3. Chapter 5 - Machine learning diagnostic analysis of human identified qualitative features

Aim: To develop machine learning models that can accurately identify oral potentially malignant disorders and oral squamous cell carcinoma based on human observed qualitative features observed on in vivo captured fluorescence confocal endomicroscopy images.

Null Hypothesis (H_0): That machine learning using human observed qualitative features of in vivo captured fluorescence confocal endomicroscopy images cannot accurately identify instances of oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That machine learning using human observed qualitative features of in vivo captured fluorescence confocal endomicroscopy images can accurately identify instances of oral potentially malignant disorders and oral squamous cell carcinoma.

1.10.4. Chapter 6 - Machine learning diagnostic analysis of quantitative feature extraction

Aim: To develop, evaluate, and compare ML models for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images.

Null Hypothesis (H_0): That ML models developed for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from

fluorescence human in vivo captured confocal endomicroscopy images cannot accurately detect oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That ML models developed for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images can accurately detect oral potentially malignant disorders and oral squamous cell carcinoma.

1.10.5. Chapter 7 - Convolutional neural network diagnostic classification

Aim: To develop and test deep learning convolutional neural network (CNN) models with on fluorescence in vivo confocal microscopy images of the oral mucosa for the rapid and accurate detection of oral potentially malignant disorders and oral squamous cell carcinoma.

Null Hypothesis (H_0): That deep learning CNN models cannot accurately and rapidly detect oral potentially malignant disorders and oral squamous cell carcinoma in fluorescence in vivo confocal microscopy images of the oral mucosa.

Alternative Hypothesis (H_1): That deep learning CNN models can accurately and rapidly detect oral potentially malignant disorders and oral squamous cell carcinoma in fluorescence in vivo confocal microscopy images of the oral mucosa.

1.10.6. Chapter 8 - Deep learning diagnostic classification in a pre-clinical murine model of oral carcinogenesis

Aim: To develop and evaluate the performance of deep learning convolutional neural network (CNN) models for the rapid and accurate classification of oral

epithelial dysplasia (OED) and oral squamous cell carcinoma using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

Null Hypothesis (H_0): That a trained deep learning model cannot rapidly and accurately detect OED and oral squamous cell carcinoma using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

Alternative Hypothesis (H_1): That a trained deep learning model can rapidly and accurately detect OED and oral squamous cell carcinoma using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

2. SYSTEMATIC REVIEW ON CONFOCAL MICROSCOPY IN THE DIAGNOSIS OF ORAL CANCER AND ORAL POTENTIALLY MALIGNANT DISORDERS

This chapter presents a systematic review of the literature involving the use of confocal microscopy for the detection of oral squamous cell carcinoma (OSCC) and oral potentially malignant disorders (OPMD).

While confocal microscopy has been proposed as a promising non-invasive technique for assessing OSCC and OPMDs, there is a wide range of device designs, functions, and applications. Further, diagnostic accuracy, clinical utility, and practical integration into routine workflows were approached diversely. Individual studies have reported encouraging results, but the evidence is dispersed and methodologically diverse, with inconsistencies in study design, imaging protocols, and diagnostic criteria.

The confocal microscopy analysis results across the studies encompassed qualitative and quantitative approaches. The extent of machine learning use across the literature was of interest. As such, it was difficult to determine the extent to which confocal microscopy could reliably contribute to early detection or monitoring of OSCC and OPMDs.

A systematic review was required to critically evaluate and synthesise the available literature, assess the quality and consistency of the evidence, and determine whether current data supported its clinical use. This review was carried out to help identify gaps in the existing research and inform the design of future studies in this dissertation aimed at assessing protocols utilising confocal microscopy in the diagnosis of oral cancer and oral potentially malignant disorders.

The chapter is comprised of a manuscript which was published in Oral Diseases (Wiley) on 23rd June 2022. The authors' accepted version is presented, however the font, referencing and numbering of the tables and figures have been modified to align with the formatting structure of this thesis.

Ramani, R. S., Tan, I., Bussau, L., Angel, C. M., McCullough, M., & Yap, T. (2022). Confocal microscopy in oral cancer and oral potentially malignant disorders: A systematic review. Oral Diseases, 00,1-13. https://doi.org/10.1111/odi.14291

In this publication I was the primary and corresponding author, and my contributions involved conceptualisation, data curation, formal analysis, investigation, methodology, visualization, writing the original draft, reviewing, and editing.

The contributions of my co-authors were as follows: Ivy Tan: conceptualisation, data curation, methodology; Lindsay Bussau: reviewing, and editing; Christopher Angel: reviewing, and editing; Michael McCullough: supervision,

reviewing, and editing; Tami Yap: conceptualisation, data curation, formal analysis, methodology, supervision, reviewing, and editing.

I would like to acknowledge the Robert and Gillian Cook Family Award for supporting this work and open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

2.1. Abstract

Objective: To systematically identify and summarise current research on the utility of confocal microscopy in oral squamous cell carcinoma and oral epithelial dysplasia in oral potentially malignant disorders.

Methods: Databases Medline, Embase, Evidence Based Medicine and Web of Science were searched with articles screened and included if their primary objective was the use of a confocal microscope in diagnosis of oral cancer or epithelial dysplasia, *in vivo* or *ex vivo*.

Results and Discussion: Twenty-eight relevant studies were identified of which 21 studies included oral squamous cell carcinoma specimens. Fifteen studies included *in vivo* use. The studies included both qualitative and fluorescence confocal microscope and reflectance confocal microscope analysis along with quantitative analysis of carcinoma and dysplasia. Thirteen studies reported the predictive value of their confocal device in the diagnosis of dysplasia and carcinoma. The quantitative software-based studies show promise in objectifying the diagnostic process for identifying abnormalities within the microstructure of the oral mucosa.

Conclusions: There was heterogeneity in the criteria for diagnosis of dysplasia and oral squamous cell carcinoma with experience levels of assessors impacting method efficacy. Both qualitative and quantitative confocal assessment methodologies have been explored, the latter highlighting the potential of future machine-augmented diagnostic precision.

Keywords: Oral cancer; confocal microscopy; dysplasia; in vivo microscopy

2.2. Conflicts of interest

No funding was received for completion of this review. The authors disclose that Optiscan Pty Ltd, together with the Melbourne Dental School, was awarded funding by the Australian Government through the Medical Research Future Fund to collaborate in clinical research utilizing confocal microscopy to improve screening and early diagnosis of oral cancer.

2.3. Funding

No funding was received for the completion of this review.

2.4. Abbreviations

AC = Actinic cheilitis

EGFR = Epidermal growth factor receptor

FCM = Fluorescence confocal microscope

OED = Oral epithelial dysplasia

OLP = Oral lichen planus

OPMD = Oral potentially malignant disorders

OSMF = Oral submucous fibrosis

OSCC = Oral squamous cell carcinoma

RCM = Reflectance confocal microscope

2.5. Introduction

Oral squamous cell carcinoma (OSCC) remains the 6th most common cancer in the world with an estimated incidence of 354,864, and the estimated mortality of 177,384 for 2020 (AIHW, 2021). Early diagnosis of this condition can reduce both morbidity and mortality. Identification and precision assessment of pre-cursor lesions, collectively termed oral potentially malignant disorders (OPMD), may be key to this stage shift (McCullough et al., 2010a; Warnakulasuriya, Kujan, Aguirre-Urizar, Bagan, González-Moles, Kerr, Lodi, Mello, Monteiro, & Ogden, 2021).

The current gold standard in diagnosis of OSCC remains incisional biopsy followed by histopathological analysis (McCullough et al., 2010a). The limitations of this procedure range from processing time, costs, patient morbidity, to potential sampling errors. Thus, there is an attraction in the development of real time in vivo imaging, such as confocal microscopy (Chu, 2010). Confocal microscopy is an optical imaging method that has the potential to achieve sub-cellular resolution for in vivo imaging at a reasonably high frame rate (Joey M. Jabbour, Meagan A. Saldua, Joel N. Bixler, & Kristen C. Maitland, 2012). These microscopes are known for their sharp images owing to the placement of a 'pinhole' between the detector and the objective lens. This provides the ability to filter in light from a single small source on the sample and reject the scattering rays from the surroundings that could distort the image (Chu, 2010). As opposed to the cross-sectional orientation of histological analysis, confocal microscopy images when captured in vivo are oriented parallel to the surface of the tissue, known as ‘en face’ (Joey M. Jabbour et al., 2012). The hand-held miniature version of this microscope that could be used intraorally, the core of the optical fibre assumes the role of the 'pinhole'. Major challenges to this hand-held technology are high resolution, fast frame rates and an adequate imaging depth for diagnosis (T. D. Wang et al., 2003).

Confocal imaging has shown the potential to streamline the process of diagnosis of oral mucosal lesions. A review by Volgger et al. (2013) opened discussion regarding the potential of confocal microscopy in head & neck disorders (Volgger, Conderman, & Betz, 2013a). This was followed by a review by Lucchese et al. (2016) which recognised the potential of reflectance confocal imaging in intra oral diagnosis (Lucchese et al., 2016a). Maher et al. (2016) summarised the in vivo applications of confocal microscopes for oral mucosal pathologies. They recognised that the quality of evidence in the studies they analysed was poor and the data was mainly limited to small descriptive studies (N. G. Maher et al., 2016). The current review aims to systematically identify and summarise current research on the utility of confocal microscopy for the

description and diagnosis of oral squamous cell carcinoma and oral epithelial dysplasia in oral potentially malignant disorders

2.6. Methodology

The current review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement 2020 (Page et al., 2021) (Figure 1). Included articles for the current review were original research studies which utilised confocal microscopy in diagnosing OSCC or OPMDs in vivo or ex vivo as defined by standard histopathology. The article search was conducted among the databases of Medline, Embase, Evidence Based Medicine and Web of Science to include all articles published up to August 22nd, 2020. An updated search was performed on 6th October 2021, with the same search terms using the same databases. The articles found in the search results were catalogued in Endnote® (Version X9.3.3, Clarivate Analytics). Search terms are listed in Appendix 1.

English-language studies with any model of confocal microscope were included if performed on either human or animal subjects. Any non-English studies or studies involving the use of a confocal microscope on oral cancer cell lines were excluded. Review articles, withdrawn/retracted studies, commentaries, and unpublished articles were also excluded. Eligibility assessment was done by two reviewers (I.T. and R.S.R.), who conducted a completed blinded screening by title, abstract and full text. The process of review was mediated by a third reviewer (T.Y.). Disagreements between reviewers were resolved by consensus.

The risk of bias assessment was done by two reviewers (R.S.R. & T.Y.) using the National Heart, Lung, and Blood Institute (NHLBI) Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies (2021) (Supplementary table) (NHLBI, 2021). One author (R.S.R.) extracted the data from included studies, and the second author (T.Y.) checked the extracted data. Disagreements were resolved by discussion between the two review authors (R.S.R., T.Y.). The data collected from the studies was separated into the following categories: year, nationality, number of subjects, specimen type, contrast agent, disorders studied, tissue sample, assessment methodology, qualitative criteria assessed, quantitative criteria assessed, tool for quantitative analysis, diagnosis, sensitivity, and specificity. Within the included studies, the OSCC/OPMD confocal microscopy assessment methodology have been explored with the sensitivity and specificity of diagnosis expressed and compared where possible. The methodology for this current systematic review has been registered on PROSPERO (National Institute for Health Research) with ID no. CRD42021279967. Meta analysis for the included studies was not conducted due to the heterogeneity of data.

2.7. Results and Discussion

A total of 28 relevant studies were identified for inclusion in the review (Supplementary figure). The search of Medline, Embase, Evidence Based Medicine and Web of Science databases provided a total of 4129 citations.

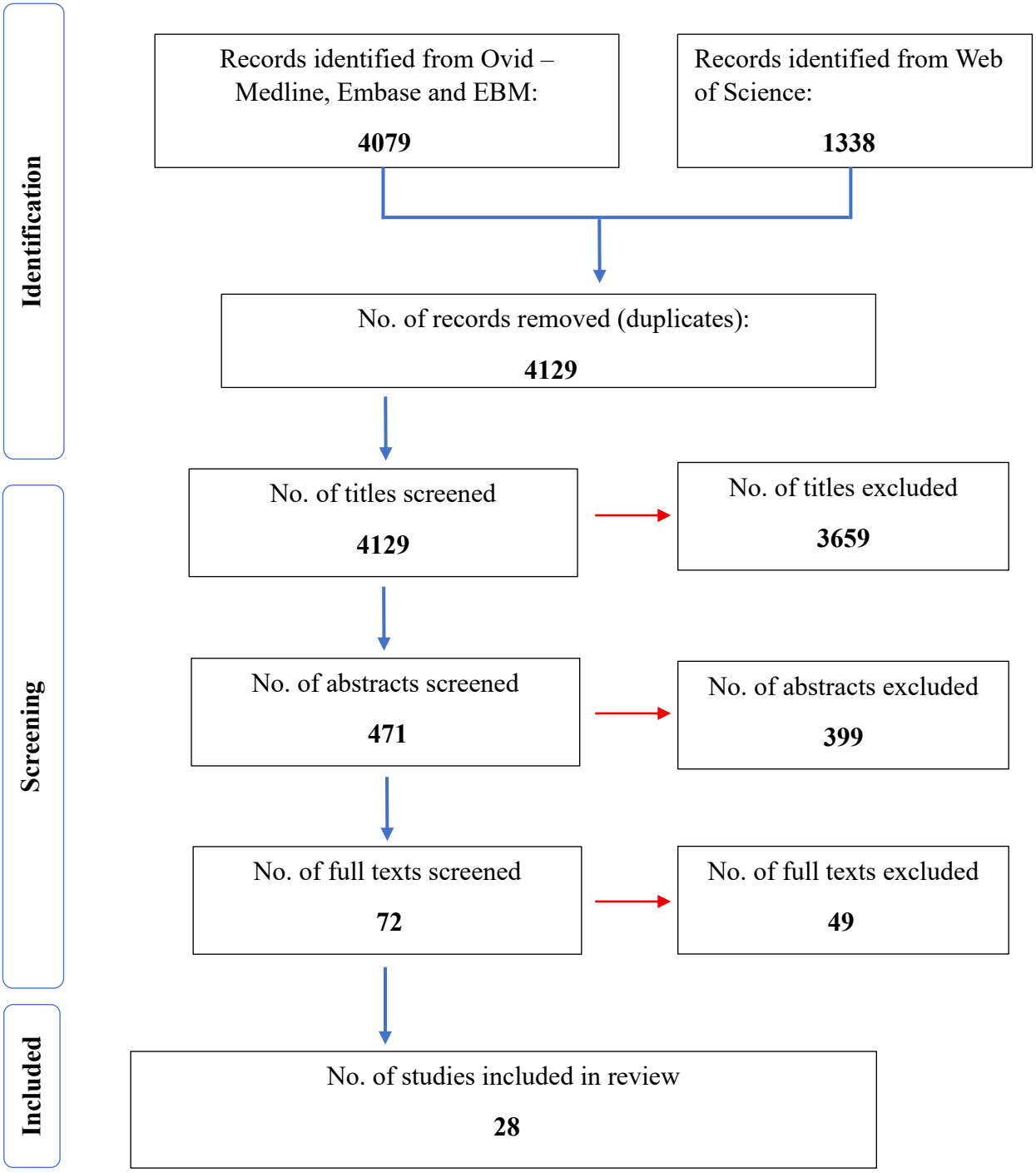


Figure 2.1. PRISMA flowchart for article search and inclusion

Of these, 3659 studies were discarded as the titles did not match the inclusion criteria. Of the remaining 471 citations, 399 were rejected because the abstracts clearly did not meet our inclusion criteria (Figure 2.1.).

Out of the 28 studies which underwent risk of bias assessment using the NHLBI Quality Assessment Tool, only 2 studies were rated "Good", 13 were rated "Fair", and the remaining 13 were rated "Poor". A major point which all studies in this review failed to follow was a justification of their sample size with a power analysis (Table 2.1.). The scores were marked based on 9 eligible questions within the NIH quality assessment tool for observational cohort and cross-sectional studies. The grading of scores was done in the following manner: 0-3 = Poor, 4-6 = Fair, and 7-9 = Good. The full text of the remaining 72 citations was examined in more detail out of which 28 articles were included in our review. An overview of the included studies is presented in Table 2.2.

Table 2.1. Studies were assessed using the NIH quality assessment tool for observational cohort and cross-sectional studies

Study	Was the research question or objective in this paper clearly stated?	Was the study population clearly specified and defined?	Was the participation rate of eligible persons at least 50%?	Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants?	Was a sample size justification, power description, or variance and effect estimates provided?	For the analyses in this paper, were the exposure(s) of interest measured prior to the outcome(s) being measured?	Was the timeframe sufficient so that one could reasonably expect to see an association between exposure and outcome?	Did the study examine different levels of the exposure as related to the outcome?	Were the exposure measures (independent variables) clearly defined, valid, reliable, and implemented consistently?	Was the exposure(s) assessed more than once over time?	Were the outcome measures (dependent variables) clearly defined, valid, reliable, and implemented consistently?	Were the outcome assessors blinded to the exposure status of participants?	Was loss to follow-up after baseline 20% or less?	Were key potential confounding variables measured and adjusted statistically for their impact on the relationship between exposure(s) and outcome(s)?	Score ('Yes' responses)	Summary of Quality
Alessi et al. (2013)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	2	P
Anuthama et al. (2010)	√	√	NA	NR	X	NA	NA	NA	√	X	√	NR	NA	√	5	F
Carlson et al. (2007)	√	X	NA	NR	X	NA	NA	NA	√	X	√	NR	NA	√	4	F
Clark et al. (2002)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	2	P
Clark et al. (2003)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	2	P
Contaldo et al. (2019)	√	√	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	3	P
Contaldo et al. (2020)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	√	NA	√	4	F
Dittberner et al. (2016)	√	√	NA	√	X	NA	NA	NA	√	X	√	CD	NA	√	6	F
Dittberner et al. (2021)	√	√	NA	√	X	NA	NA	NA	√	X	√	√	NA	√	7	G
El Hallani et al. (2013)	√	√	NA	NR	X	NA	NA	NA	√	X	√	NR	NA	X	4	F
Farahati et al. (2010)	√	√	NA	√	X	NA	NA	NA	√	X	CD	NR	NA	√	5	F

Haxel et al. (2010)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	2	P
Hellbust et al. (2013)	√	√	NA	√	X	NA	NA	NA	√	X	CD	NR	NA	X	4	F
Linxweiler et al. (2016)	√	X	NA	NR	X	NA	NA	NA	√	X	CD	NR	NA	X	2	P
Lupu et al. (2018)	√	X	NA	X	X	NA	NA	NA	√	X	X	NR	NA	X	1	P
Lupu et al. (2020)	√	X	NA	X	X	NA	NA	NA	√	X	X	NR	NA	X	3	P
Maitland et al. (2008)	√	X	NA	NR	X	NA	NA	NA	√	X	X	NR	NA	X	2	P
Moore et al. (2016)	√	X	NA	X	X	NA	NA	NA	√	X	X	NR	NA	√	3	P
Nathan et al. (2014)	√	√	NA	X	X	NA	NA	NA	√	X	X	√	NA	√	5	F
Oetter et al. (2016)	√	√	NA	√	X	NA	NA	NA	√	X	√	√	NA	√	7	G
Peng et al. (2020)	√	√	NA	√	X	NA	NA	NA	√	X	X	X	NA	√	5	F
Pogorzelski et al. (2012)	√	X	NA	√	X	NA	NA	NA	√	X	X	X	NA	X	3	P
Shavlokhova et al. (2020)	√	√	NA	√	X	NA	NA	NA	√	X	X	√	NA	√	6	F
Shavlokhova et al. (2021a)	√	√	NA	√	X	NA	NA	NA	√	X	X	X	NA	√	5	F
Shavlokhova et al. (2021b)	√	√	NA	√	X	NA	NA	NA	√	X	X	X	NA	X	4	F
Shinohara et al. (2020)	√	X	NA	NR	X	NA	NA	NA	√	X	X	NR	NA	X	2	P
Sievert et al. (2021)	√	X	NA	√	X	NA	NA	NA	√	X	X	√	NA	√	5	F
Ulrich et al. (2011)	√	X	NA	NR	X	NA	NA	NA	√	X	X	√	NA	X	2	P

√ = Yes, X = No, NA = Not applicable, NR = Not reported, CD = Cannot determine, P = Poor, F = Fair, and G = Good

Table 2.2. Summary of included studies

Author and year	Nationality	Subjects	Number of subjects	Specimen type	Contrast agent	Disorders studied
Alessi et al. (2013)	Brazil	Human	25	In vivo	None	OLP
Anuthama et al. (2010)	India	Human	25	Ex vivo	6% Acetic acid	Betel chewer's mucosa, Leukoplakia, OSMF and OSCC
Carlson et al. (2007)	U.S.A.	Human	14	Ex vivo	Anti-EGFR antibody with fluorescent dye and 6% Acetic acid	OED and OSCC
Clark et al. (2002)	U.S.A.	Human	6	Ex vivo	6% Acetic acid	OSCC
Clark et al. (2003)	U.S.A.	Human	17	Ex vivo	6% Acetic acid	OED and OSCC
Contaldo et al. (2019)	Italy	Human	12	In vivo	None	OLP
Contaldo et al. (2020)	Italy	Human	21	In vivo	None	Leukoplakia and OSCC
Dittberner et al. (2016)	Germany	Human	12	In vivo	Fluorescein	OSCC
Dittberner et al. (2021)	Germany	Human	13	In vivo	Fluorescein	OSCC
El Hallani et al. (2013)	Canada	Human	13	Ex vivo	Acriflavine & Cresyl Violet	OED
Farahati et al. (2010)	Germany	Animal	58	In vivo	None	OED and carcinoma-in-situ
Haxel et al. (2010)	Germany	Human	5	Both	Fluorescein & Acriflavine	OSCC
Hellbust et al. (2013)	U.S.A.	Animal	25	Ex vivo	Proflavin	OED and OSCC
Linxweiler et al. (2016)	Germany	Human	185	Ex vivo	Acriflavine	OSCC
Lupu et al. (2018)	Romania	Human	2	In vivo	None	AC and OSCC
Lupu et al. (2020)	Romania	Human	12	In vivo	None	AC and OSCC
Maitland et al. (2008)	U.S.A.	Human	8	In vivo	Apple cider vinegar	OED and OSCC

Moore et al. (2016)	U.S.A.	Human	21	In vivo	Fluorescein	OED and OSCC
Nathan et al. (2014)	U.S.A.	Human	21	In vivo	Fluorescein	Leukoplakia and OSCC
Oetter et al. (2016)	Germany	Human	95	In vivo	Fluorescein	OSCC
Peng et al. (2020)	China	Human	47	In vivo	None	OLP
Pogorzelski et al. (2012)	Germany	Human	15	Both	Fluorescein & Acriflavine	OSCC
Shavlokhova et al. (2020)	Germany	Human	70	Ex vivo	Acridine orange	OSCC
Shavlokhova et al. (2021b)	Germany	Human	22	Ex vivo	Acridine orange	OED
Shavlokhova et al. (2021a)	Germany	Human	35	Ex vivo	Acridine orange	OSCC
Shinohara et al. (2020)	Japan	Human	10	Ex vivo	Acriflavine & Edible food dye	OSCC
Sievert et al. (2021)	Germany	Human	5	In vivo	Fluorescein	OSCC
Ulrich et al. (2011)	Germany	Human	10	In vivo	None	AC & OED

OLP = Oral lichen planus, OSCC = Oral squamous cell carcinoma, OED = Oral epithelial dysplasia, AC = Actinic cheilitis, OSMF = Oral submucous fibrosis, EGFR = Epidermal Growth Factor Receptor

The assessment methodology employed by the included studies with the use of confocal microscopy for description and diagnosis of OPMDs and OSCC are presented in Table 2.3. Twenty-one studies utilised qualitative assessment methods to this end and the remaining seven used quantitative assessment methods (Table 2.3.).

Table 2.3. Confocal assessment methodology

Author and year	Tissue sample	Assessment methodology	Qualitative criteria assessed	Quantitative criteria assessed	Tool for Quantitative analysis	Diagnosis	Sensitivity	Specificity
Alessi et al. (2013)	In vivo	Histopathologic correlation	Integrity of all epithelium layers, presence of epithelial clefts, acantholytic cells, inflammatory cells	N.A.	N.A.	OLP	-	-

			within the epithelium, dendritic cells, and dilated blood vessels					
Anuthama et al. (2010)	Ex vivo	Image analysis software and histopathologic correlation	Keratin deposition, cell shape, nuclei, intensity of connective tissue stroma, and epithelial thickness	Keratin thickness, cell density, and nuclear density along with the mean intensity of different components of the epithelium	Leica image analysis software (LIAS)	Betel Chewer's mucosa, Leukoplakia, OSMF, and OSCC	-	-
Carlson et al. (2007)	Ex vivo	Image analysis software and histopathologic correlation	Epithelium architecture, cellular and nuclear density	Fluorescence labelling intensity (FLI) and nuclear cytoplasmic ratio (N/C ratio)	MATLAB software	OSCC and dysplasia	FLI - 1, N/C - 0.9	FLI - 0.86, N/C - 1
Clark et al. (2002)	Ex vivo	Identification of confocal features	Cellular sizes, cell outlines, nuclei sizes and nuclei shapes	N.A.	N.A.	OSCC	-	-
Clark et al. (2003)	Ex vivo	Histopathological correlation	Cellular sizes, cell density, nuclear intensity, and signs of inflammation.	N.A.	N.A.	OSCC	-	-
Contaldo et al. (2019)	In vivo	Histopathologic correlation	Type of keratosis, size of keratinocytes, nuclear size and	N.A.	N.A.	OLP	-	-

			shape, acanthosis and spongiosis, necrotic keratinocytes, connective tissue papilla and inflammatory cells					
Contaldo et al. (2020)	In vivo	Histopathological correlation	Thickness of epithelium, epithelial architecture, keratosis, size and shape of keratinocytes, and keratin pearls	N.A.	N.A.	OSCC and dysplasia	1	0.93
Dittberner et al. (2016)	In vivo	Image analysis software	N.A.	Epithelial cell sizes by automated cell border segmentation and distance map between cell borders	Image analysis algorithm	OSCC	0.72	0.85
Dittberner et al. (2021)	In vivo	2 assessors & histopathological correlation	DOC score (developed by Oetter et al. 2016): tissue architecture, cell morphology, fluorescein leakage	N.A.	N.A.	OSCC	0.87	0.80
El Hallani et al. (2013)	Ex vivo	Image analysis software used and correlation with histology	Cellular spacing, uniformity of size and shape of cells	Calculation of cellular density	Getafics image analyser software	OED	-	-

Farahati et al. (2010)	In vivo	2 assessors & histopathological correlation	Cellular size, nuclear size and keratotic areas in epithelium	N.A.	N.A.	OED	0.73	0.88
Haxel et al. (2010)	Both	Histopathological correlation	Cellular structure, epithelial architectural features, connective tissue and neoangiogenesis.	N.A.	N.A.	OSCC	-	-
Hellebust et al. (2013)	Ex vivo	Histopathological correlation	Changes in keratin structure, nuclear crowding, and nuclear enlargement	N.A.	N.A.	OSCC	-	-
Linxweiler et al. (2016)	Ex vivo	12 assessors (4 surgeons, 4 pathologists, 4 laymen) & histopathological correlation	Tumour border and tumour localization	N.A.	N.A.	OSCC	-	-
Lupu et al. (2018)	In vivo	Histopathological correlation	Epithelial architecture, cellular shape, intercellular spaces, appearance of blood vessels and inflammatory cells	N.A.	N.A.	AC and OSCC	-	-
Lupu et al. (2020)	In vivo	Image analysis software	Ulceration, keratosis, cellular architecture pattern, dendritic cells, solar elastosis, dermal inflammatory cells, and tumour	The mean diameter of blood vessels and the mean number of blood vessels	ImageJ	AC and OSCC	-	-

			nests in dermis					
Maitland et al. (2008)	In vivo	Histopathological correlation	Dispersed nuclei, dense nuclei and disordered tissue structure	N.A.	N.A.	OSCC	-	-
Moore et al. (2016)	In vivo	8 assessors (7 surgeons, 1 pathologist) & histopathological correlation	Epithelial architecture, cellular shape	N.A.	N.A.	OED & OSCC	OED = 0.80, OSCC = 0.86	OED = 1, OSCC = 1
Nathan et al. (2014)	In vivo	4 assessors (3 surgeons, 1 pathologist) & histopathological correlation	Cellular shape, epithelial architecture, and vasculature	N.A.	N.A.	OED & OSCC	OED = 0.8, OSCC = 0.86	OED = 1, OSCC = 1
Oetter et al. (2016)	In vivo	6 assessors (3 expert, 3 non-expert) using developed scoring system (DOC-Score) & histopathological correlation	Homogeneity of tissue architecture, intercellular gaps, cell morphology, fluorescence leakage and vessels.	N.A.	N.A.	OSCC	0.95	0.88
Peng et al. (2020)	In vivo	Histopathologic correlation	Parakeratosis, acanthosis, liquefaction degeneration, inflammatory cell infiltration and dilated blood vessels	N.A.	N.A.	OLP	0.50	0.79
Pogorzelski et al. (2012)	Both	Histopathological correlation	Epithelial architecture, cellular size and shape and characterist	N.A.	N.A.	OSCC	-	-

			ics of blood vessels					
Shavlokhova et al. (2020)	Ex vivo	3 assessors (2 surgeons, 1 pathologist) & histopathological correlation	Cellular shape, size, nuclear atypia or pleomorphism, and increased nuclear density	N.A.	N.A.	OSCC	0.91	0.75
Shavlokhova et al. (2021a)	Ex vivo	Histopathologic correlation	Asymmetrical epithelial stratification, increased mitotic figures, dyskeratosis, drop-shaped rete pegs, keratin pearls, nuclear pleomorphism, cellular pleomorphism, increase in nuclear–cytoplasmic ratio, prominent nucleoli, and hyperchromatism	N.A.	N.A.	OED	0.96	0.92
Shavlokhova et al. (2021b)	Ex vivo	Histopathologic correlation	Cellular pleomorphism, nuclear hyperchromatism, prominent nucleoli, increase in nuclear cytoplasmic ratio, loss of cellular adhesion, keratinisation	N.A.	N.A.	OSCC	-	-

Shinohara et al. (2020)	Ex vivo	Measured autofluorescence intensity and histopathologic correlation	Nuclear size, shapes, and patterns	To detect changes in reflectance intensity of nuclei and cytoplasm between the normal and SCC tissue	Spectrometer	OSCC	-	-
Sievert et al. (2021)	In vivo	4 assessors (3 surgeons, 1 pathologist) & histopathological correlation	Cellular size, shape, cytoplasmic membrane, blood vessel architecture	N.A.	N.A.	OSCC	0.90	0.78
Ulrich et al. (2011)	In vivo	Defined scoring criteria & histopathologic correlation	Disruption of stratum corneum, hyperkeratotic scale, parakeratosis, cellular pattern, solar elastosis, blood vessel dilation, and inflammation	N.A.	N.A.	AC	-	-

N.A. = Not available, AC = Actinic cheilitis, OSCC = Oral squamous cell carcinoma, OLP = Oral lichen planus, OSMF =Oral submucous fibrosis, OED = Oral epithelial dysplasia, FLI = Fluorescence labelling intensity, N/C = Nuclear cytoplasmic ratio

2.7.1. In vivo OSCC studies

Eight studies in this review investigated in vivo OSCC samples. Seven of these used qualitative image analyses and the remaining 1 used quantitative analysis software (Table 2.3.). Of these 7 qualitative analysis studies, Maitland et al. (2008) and Contaldo et al. (2020) performed a direct identification of confocal microscopy features and correlated them with histopathology. Both these studies avoided using fluorescence biomarkers and Contaldo et al. (2020) reported a relatively high sensitivity and specificity (Table 2.3.). The

characteristic RCM signs of OSCC in these studies were cellular and architectural disarray, keratin pearls, cellular pleomorphism and increased nuclear cytoplasmic ratios (Contaldo et al., 2020; Maitland et al., 2008). However, the limitations of RCM such as poor depth penetration, strong backscatter of light in highly keratinised tissue, and diffusion of dense inflammatory infiltrates should be kept in mind (Contaldo et al., 2020).

The remaining 5 qualitative in vivo OSCC studies utilized multiple assessors of confocal microscopy images and correlated their results with histopathology (Table 2.3.). These studies generally found good agreement between their assessors, but examiners experienced with the technology recorded higher diagnostic accuracy. The highest sensitivity and specificity of OSCC identification in these multiple assessor studies belonged to Oetter et al. (2016) (Table 2.3.). The authors of this study defined their own OSCC confocal diagnostic criteria termed as the DOC-Score. This score focused on homogeneity of tissue architecture, intercellular gaps, cell morphology, fluorescence leakage and integrity of blood vessels (Oetter et al., 2016). The efficacy of this scoring system was tested when Dittberner et al. (2021) used it for their in vivo OSCC study and received a relatively high sensitivity and specificity score (Table 2.3.) (Dittberner et al., 2021). Despite its relative success in these studies, the DOC score leans towards direct comparisons with the landmarks seen in histopathology. Scoring criteria designed specifically for landmarks observed in the en face orientation of confocal microscopy with different contrast agents must be developed. Despite not using a specific scoring criteria, Sievert et al. (2021) also utilized fluorescein in their in vivo OSCC study and on comparison with their histopathology results obtained a high sensitivity and specificity of, respectively, 89.9% and 78.6%. (Sievert et al., 2021). However, region of interest selection in qualitative analysis could be subject to investigator bias.

One in vivo OSCC study conducted by Dittberner et al. (2016) focused on quantitative analysis of OSCC tissue using an image analysis algorithm which could draw and superimpose cell borders over the FCM images of in vivo oral mucosa and calculate the distance between the cell borders. This algorithm learned the patterns of cell sizes in normal and malignant tissue using training image data from 10 subjects (4 cancer patients & 6 control subjects). The authors used 2 subjects as test data with relatively high accuracy (Table 2.3.) (Dittberner et al., 2016b). Although cell size seems to be an appropriate measure there are several oral mucosal abnormalities, injuries, and tissue repair mechanisms that might influence variation in cell size which could provide false positives.

2.7.2. Ex vivo OSCC studies

Nine studies in this review examined ex vivo OSCC tissue. Six of these used qualitative analysis and the remaining 3 used quantitative techniques (Table 2.3.). Out of these 6 qualitative studies, 3 conducted a direct correlation of confocal results and histopathology, 2 used multiple assessors and the study by Clark et al. (2002) conducted a simple description of the OSCC tissue they observed with a confocal microscope (Table 2.3.) (A. Clark et al., 2002).

Clark et al. (2003) in their RCM ex vivo OSCC study simply described the different patterns they observed in their subject with histopathological confirmed OSCC (A. L. Clark et al., 2003). These descriptions while helpful, do not aid in defining diagnostic criteria due to a lack of controls. Hellebust et al. (2013) in their FCM ex vivo OSCC study found a reasonable degree of overlap between histopathology and confocal microscopy. The authors used the biomarker proflavine on ex vivo specimens from a mouse model of chemically induced tongue carcinogenesis with a 91% identification accuracy when combined with wide-field imaging (Hellebust et al., 2013). However, Shavlokhova et al. (2021a) in their FCM study of OSCC found no clear correlation between most histological features of OSCC in confocal images. They could only observe cellular pleomorphism predominantly in poorly differentiated tumours and keratinization mostly in well differentiated OSCC (Shavlokhova, Flechtenmacher, Sandhu, Vollmer, Hoffmann, et al., 2021). It could be misleading to observe features in en face images and compare them to those seen in the cross-sectional orientation of ex vivo confocal microscopy and histopathology.

Linxweiler et al. (2016) and Shavlokhova et al. (2020) conducted their ex vivo OSCC studies with assessors of different levels of confocal microscopy experience. Linxweiler et al. (2016) provided a basic explanation to 4 head and neck surgeons, 4 pathologists, and 4 laymen to judge the FCM characteristics of malignancy. Correct identification of tumour demarcation had a success rate of 97% in pathologists, 85% in head and neck surgeons, and 70% in laymen. It was interesting to note the relatively high success rate of laymen following a basic explanation of confocal image analysis for malignancies (Linxweiler et al., 2016a). Shavlokhova et al. (2020) involved a pathologist, a maxillofacial surgeon (novice physician), and a maxillofacial surgeon trained in in vivo RCM (expert physician) in their study. Using the diagnostic criteria of nuclear atypia, pleomorphism and nuclear density, the pathologist (sensitivity = 0.99, specificity = 0.95) showed higher accuracy compared to the expert physician (sensitivity = 0.96, specificity = 0.66) and novice physician (sensitivity = 0.98, specificity = 0.57). The authors acknowledged that a lack of specific standardised OSCC criteria made it difficult to determine an assessment

protocol (Shavlokhova et al., 2020). These studies note an increased diagnostic accuracy with greater confocal microscopy experience and training.

Carlson et al. (2007), Anuthama et al. (2010), and Shinohara et al. (2020) used quantitative techniques in their ex vivo OSCC diagnosis. Carlson et al. (2007) used MATLAB (The MathWorks, Inc, Natick, MA) to determine the fluorescence labelling intensity (FLI) of anti-epidermal growth factor receptor (EGFR) antibody contrast expression in the stained tissue samples. Differentiating moderate dysplasia/carcinoma VS normal/mild dysplasia, produced a sensitivity of 1 which was the highest among all quantitative assessment methodologies and among all the studies in the review. The carcinoma samples showed significantly greater FLI values (Wilcoxon rank sum test p value < 0.01) than the normal/mild dysplasia samples. When comparing FLI values to separate carcinoma/dysplasia VS normal mucosa the sensitivity and specificity were 0.77 and 0.84, respectively. Nuclear-cytoplasmic (N/C) ratio from the RCM images in this study was calculated using MATLAB. Differentiating severe dysplasia/carcinoma VS normal/mildly dysplastic tissue using N/C ratio, produced a sensitivity and specificity of 0.90 and 1, respectively (Carlson, Gillenwater, Williams, El-Naggar, & Richards-Kortum, 2007a). FLI and N/C ratio calculations appear promising for the diagnosis of OED and OSCC, due to the quantifiable visual feedback received.

While Carlson et al. (2007) observed exogenous fluorescence intensity, Shinohara et al. (2020) measured autofluorescence in OSCC by using a real-time spectrometer to quantify its intensity and spectrum. They found that normal mucosa emitted higher autofluorescence than malignant tissues at 473 nm excitation. The authors found it difficult to differentiate between normal tissue and carcinoma based on autofluorescence alone due to irregular reflections on their probe or on the fibre bundle connected to the microscope (Shinohara et al., 2020). This lack of imaging consistency with autofluorescence could warrant the use of exogenous fluorescence instead of autofluorescence.

Anuthama et al. (2010) conducted their quantitative RCM analysis by including samples of betel chewers mucosa (BCM), leukoplakia, oral submucous fibrosis (OSMF) and OSCC. They used Leica image analysis software (Leica Microsystems, Germany) to measure nuclear density in terms of the wavelength of reflected light depicted in nanometres (nm). The authors noted that OSCC samples showed a higher mean nuclear density value (245.91 nm) compared to OSMF (174.53 nm), leukoplakia (154.22 nm), BCM (119.70 nm), and normal controls (94.43 nm). This high nuclear density of OSCC was attributed by the authors to active division and proliferation of malignant cells (Anuthama et al., 2010). Another possible cause for an increase in nuclear density is speculated to be chromatin compaction (Huisman et al., 2005). Studying nuclear density in

dysplasia and OSCC using confocal microscopy could be useful for determining prognosis of oral lesions.

2.7.3. In vivo OPMD studies

Seven studies in this review studied in vivo OPMD specimens with only one using a quantitative assessment method (Table 2.3.).

Alessi et al. (2013), Contaldo et al. (2019) and Peng et al. (2020) studied in vivo oral lichen planus (OLP)(Alessi, Nico, Fernandes, & Lourenço, 2013; Contaldo, Di Stasio, Petruzzi, Serpico, & Lucchese, 2019; Peng, Shen, Zhou, & Wang, 2020). Alessi et al. (2013) studied RCM features seen in human desquamative gingivitis for the in vivo diagnosis of OLP. They found inflammatory cells adjacent to the basal keratinocytes around the submucosal papillae along with dark intercellular spaces and inflammatory cell infiltrate in the epithelium. Additionally present were bright isolated cells that appeared larger than surrounding keratinocytes without a visible nucleus at the spinous layer level, which corresponded to civatte bodies found in histological sections (Alessi et al., 2013). Contaldo et al. (2019) took these RCM OLP observations further by correlating them to clinical variants of OLP such as reticular, atrophic-erosive, and mixed pattern in 31 OLP lesions across 12 patients. The reticular lesions presented parakeratosis along with large, polygonal keratinocytes filled by granules surrounding the nucleus in a dark halo like pattern. All the OLP samples also showed cloudy polygonal single cells representing necrotic keratinocytes. Inflammation and obscured connective tissue papillae were highlighted in the atrophic-erosive lesions due to the lack of hyperkeratosis (Contaldo et al., 2019). Peng et al. (2020) also noticed ring-like bright structures at the epithelial-connective tissue junction which were attributed to keratinocyte necrosis in their RCM OLP lesions. The authors noted dark lumen structures with bright round cells inside these structures which corresponded to dilated vessels with inflammatory cell infiltration (Peng et al., 2020). The RCM features of OLP seen by these three studies were consistent with the ones reported in the review by Lucchese et al. (2016) (Lucchese et al., 2016a).

Farahati et al. (2010) was the only study to examine in vivo oral epithelial dysplasia (OED). The authors used RCM in the oral mucosa of 58 mice exposed to 4-nitro-quinoline 1-oxide (4-NQO) (Farahati et al., 2010). They had two examiners look at cell size, nuclear size, and keratotic areas and categorise the specimens as mild dysplasia, moderate dysplasia, severe dysplasia/carcinoma-in-situ. The inter-rater reliability for these two observers was found to be Kappa = 0.59 ($p < 0.001$) and they showed a relatively high accuracy for differentiation

between mild/moderate dysplasia and severe dysplasia/carcinoma-in-situ (Table2) (Farahati et al., 2010).

The remaining 3 in vivo OPMD studies looked specifically at actinic cheilitis (AC) and its comparison with the confocal appearance of OSCC. Ulrich et al. (2011) defined their own RCM criteria for the in vivo identification of AC which involved the disruption of stratum corneum, parakeratosis, atypical epithelial honeycomb pattern, solar elastosis, blood vessel dilation, and inflammation in the upper dermis. In this study, keratinocyte atypia and atypical honeycomb pattern represented the most valuable criteria (Ulrich et al., 2011b). Lupu et al. (2018) also used RCM for the in vivo identification of AC and noted similar findings to Ulrich et al. (2011) (Lupu et al., 2018; Ulrich et al., 2011b). Lupu et al. (2020) quantified their diagnostic method for AC and OSCC of the lip by measuring the blood vessel diameters on retrospective RCM images using an open-source software package called ImageJ. The authors found that the mean blood vessel diameter of lip OSCC (37.81 μm) was significantly larger than that of AC (19.26 μm) by 18.55 μm (p=0.006). This increase in diameter is supported by the high metabolic needs of a malignancy (Lupu, Caruntu, Boda, & Caruntu, 2020). This technique seems difficult to replicate reliably due to the varying orientations of blood vessels visible in each section and field of view.

2.7.4. Ex vivo OPMD studies

Two studies employed an ex vivo approach to examining OED specimens (Table 2.3.). El Hallani et al. (2013) used a quantitative image analyser software called Getafics to quantify the density of keratinocytes in correlation with levels of dysplasia in oral ex vivo samples. The mean keratinocyte density across the tissue samples was 5.1×10^{-4} per mm^2 in normal healthy tissue, 4.7×10^{-4} per mm^2 in hyperplastic tissue, 5.1×10^{-4} per mm^2 in mild dysplasia, 13.3×10^{-4} per mm^2 in moderate dysplasia, and 28.1×10^{-4} per mm^2 in severe dysplasia. The authors found that the mean cellular density of moderate and severe dysplasia tissues was significantly higher than the normal, hyperplasia, and mild dysplasia tissue samples (p=0.0001) (El Hallani et al., 2013). It is important to note that slice thickness of the confocal images could influence the number of keratinocytes seen in a particular field of view, which could skew these results.

Shavlokhova et al. (2021b) on the other hand studied in vivo oral leukoplakia using FCM with acridine orange. The authors looked for cytological features of dysplasia such as nuclear pleomorphism, cellular pleomorphism, nuclear–cytoplasmic ratio, prominent nucleoli, and nuclear hyperchromatism (Shavlokhova, Flechtenmacher, Sandhu, Vollmer, Vollmer, et al., 2021). The sensitivity of the ex vivo study of Shavlokhova et al. (2021b) was higher than

previous in vivo FCM OPMD studies (Table 2.3.) (Moore et al., 2016; Nathan et al., 2014; Shavlokhova, Flechtenmacher, Sandhu, Vollmer, Vollmer, et al., 2021). This indicates that OED confocal identification of ex vivo tissue samples provide a higher accuracy compared to in vivo confocal imaging and potentially the superiority of acridine orange as a biomarker over intravenous fluorescein in OED identification.

2.7.5. Studies involving in vivo and ex vivo OSCC

Two studies in this review conducted a combined in vivo and ex vivo assessment of OSCC tissue (Table 2.3.). Haxel et al. (2010) who used FCM for in vivo OSCC examination, noted that neoplastic lesions stained by fluorescein showed irregular epithelial architecture, unclear cell borders, and increased size of blood vessels. The authors followed this up by taking biopsies from these same intraoral sites and examining them ex vivo with the help of topical acriflavine. Acriflavine highlighted the prominent nuclei and increased mitotic figures (Haxel, Goetz, Kiesslich, & Gosepath, 2010). Similarly, Pogorzelski et al. (2012) used fluorescein for in vivo imaging of OSCC and witnessed highlighted extended capillaries and neoangiogenesis. Further, they conducted an ex vivo imaging with acriflavine, which highlighted the variation in nuclear sizes and subcellular details (Pogorzelski, Hanenkamp, Goetz, Kiesslich, & Gosepath, 2012a). Both these studies mentioned the limitation of topical acriflavine based confocal microscopy to the superficial 50 μm of the ex vivo tissue being examined. They also noted the difference in structures highlighted by fluorescein and acriflavine and the lack of structures visualised by autofluorescence alone (Haxel et al., 2010; Pogorzelski et al., 2012a).

2.7.6. Limitations

The exclusion of non-English studies introduces a publication bias in this review. The heterogeneity in criteria for in vivo/ex vivo diagnosis of OPMDs and OSCC make it difficult to judge the effectiveness of confocal microscopy across different examiners. The factors affecting diagnostic precision vary based on type of microscope, brand/make of microscope, contrast agent, sample size, in vivo/ex vivo tissue samples, and quantitative/qualitative confocal image analysis. Additionally, the cost of this equipment is an important consideration for widespread use. The differences in advantages of FCM and RCM need to be explored in depth. While there are a few studies that note the downsides of RCM in identifying structures clearly, more evidence is required to aid in decision making while choosing FCM or RCM in a clinical setting.

Targeting of lesion with the tip of the endoscope/confocal microscope probe could be difficult, as a stabilization of the handheld confocal microscope has been a challenge (Haxel et al., 2010). Other challenges involve artifacts due to breathing movements in the subject or mucous/blood accumulation on the optical probe of the confocal microscope (Pogorzelski et al., 2012a). Investigator experience in FCM/RCM interpretation is another factor that appears to affect qualitative findings (Linxweiler et al., 2016a; Moore et al., 2016; Oetter et al., 2016; Shavlokhova et al., 2020; Ulrich et al., 2011b). For confocal microscopy to be a viable alternative for diagnosis, extensive formal training and standardized diagnostic and hardware criteria are necessary. The software-based studies in the current review show promise in objectifying the diagnostic process for identifying abnormalities within the microstructure of the oral mucosa. Advances in image analysis software and algorithms written specifically for this purpose would aid in establishing standardised criteria for quantitative confocal analysis of oral mucosal lesions.

Very few studies reported on sensitivity and specificity for comparing confocal microscopy findings to histopathology, which would strengthen its validity in diagnosis. In a few studies that did report on accuracy of findings, ex vivo confocal diagnosis appeared to be slightly more accurate than in vivo confocal diagnosis (Moore et al., 2016; Nathan et al., 2014; Shavlokhova, Flechtenmacher, Sandhu, Vollmer, Vollmer, et al., 2021) (Table 2.3.). However, the utility of ex vivo confocal microscopy at this stage seems somewhat equivalent to histopathology and tends to share some of the same drawbacks. These ex vivo studies serve to test the functionality of confocal microscopy in identifying microscopic landmarks of cancer and OPMDs in oral mucosa, as a natural predecessor to their in vivo counterparts. The real potential for a non-invasive, real-time chairside diagnostic modality seems to lie with in vivo confocal imaging coupled with an objective image analysis software.

2.8. Conclusion

The utility of confocal microscopy has been demonstrated for tissue level diagnosis of OSCC and dysplasia in both in vivo and ex vivo specimens from the oral cavity. There is heterogeneity across the studies regarding the assessment of oral mucosa tissue. Both qualitative and quantitative confocal assessment methodologies have been explored, the latter highlighting the potential of future machine-augmented diagnostic precision.

3.METHODOLOGY

3.1. Ethics Approval

This study, compliant with the Australian Privacy Act of 1988 involved data collection from the Royal Dental Hospital Melbourne (RDHM) and Melbourne Dental School (MDS) at the University of Melbourne, Melbourne, Australia and the Walter and Eliza Hall Institute, Melbourne, Australia. All data analysis and further research was carried out at the MDS within the University of Melbourne.

This study was conducted in accordance with the Declaration of Helsinki and the data collected from human participants was under the ethics approval from the Medicine and Dentistry Human Ethics Sub-Committee under the ID 1955205 dated 04/05/2020 to the project titled “Mouth Mapping – A Systematic Assessment of the Risk of Developing Oral Cancers” and Dental Health Services Victoria (RRG: 341). Written informed consent was obtained from all patients.

All experiments involving murine data were approved by The Walter and Eliza Hall Institute Animal Ethics Committee (AEC 2018.042 approved 28/11/2018) in accordance with the Prevention of Cruelty to Animals Act (1986), the Australian National Health and Medical Research Council Code of Practice for the Care and Use of Animals for Scientific Purposes (1997), and with the ARRIVE guidelines (Percie du Sert et al., 2020).

3.2. Data collection

3.2.1. Participants

Patients attending for assessment of an oral mucosal abnormalities within the Oral Medicine Department of the Royal Dental Hospital of Melbourne were eligible for participation. Fifty-nine patients were imaged between 08/12/2020 and 13/03/2022. The patient sample had a mean age of 64.03 ± 12.66 years. There were 26 female participants with a mean age 66.92 ± 12.07 and the 33 male participants with a mean age 61.76 ± 12.82 years in this study.

Participants were recruited consecutively from patients attending the hospital's oral medicine clinic who met the study's eligibility criteria and consented to participate. Given that oral cancer and its potentially malignant disorders are relatively rare, the final sample size reflected the number of suitable cases presenting during the recruitment period rather than a predetermined target.

Three oral medicine specialists were active in collection of confocal micrograph images and had access to identifiable participant information. Research data was collected and analysed under a research ID only. Only the Project leader and coordinator had access to a locked patient ID and research ID key. All other investigators did not have access to identifiable information during or after data collection.

3.2.2. Data collection device

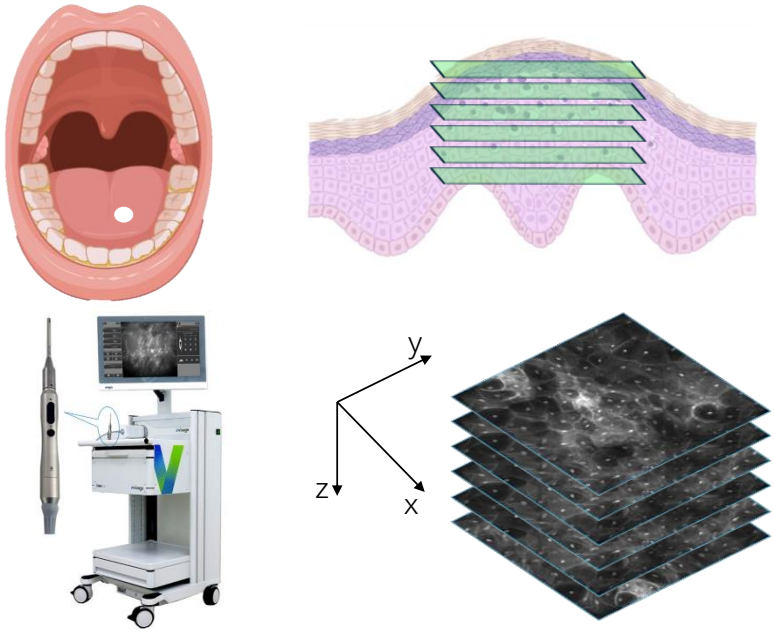


Figure 3.1. The InVivage® confocal endomicroscope (Optiscan Imaging, Australia) captures micrographs up to depths of 400µm within the oral epithelium

The in vivo confocal laser endomicroscopy (CLE) fluorescence imaging was conducted using the InVivage® (Optiscan Imaging, Victoria Australia) (Figure 3.1). This microscope was an investigational medical device at the time of the study. The device consisted of a hand-held laser scanning confocal endomicroscope that uses a single illumination and detection channel at 488 nm excitation (Table 3.1). To use the system, the probe is placed in direct contact with the tissue after the tissue has been treated with a dye. This allows the system to acquire images at different depths along the z-axis up to 400 µm. It has a field of view of 475 µm x 475 µm. The system's resolution capabilities are 0.55 µm laterally and 5.1 µm axially (Rangrez, Bussau, Ifrit, Preul, & Delaney, 2021).

Table 3.1. Technical specifications of the InVivage® (Optiscan Imaging) handheld probe (Rangrez et al., 2021)

Diameter	3.5 mm
Length	35 mm
X drive (line)	Coil and magnet
Y drive (frame)	Coil and magnet
Lateral resolution	0.55 μm
Z drive (focus)	Shape memory alloy
Axial resolution	5.1 μm
Field-of-View	475 × 475 μm
Focus range	400 μm

The system has a software interface that allows for adjustment of laser power, depth, refresh rate, capture methods and macro data that is imbedded in the image as metadata such that each image contains codes indicating the intra-oral site and research ID key where it was procured.

3.2.3. Contrast agents

Acriflavine (0.1%) and fluorescein (0.1%) were the topical contrast agents used in this study to enhance the fluorescence imaging contrast of the CLE.

The acriflavine dye used in the study was commercially compounded by a pharmacist (Como Compounding Pharmacy, South Yarra VIC 3141), to a concentration of 0.1% diluted in sterile water. Applied intra-orally by dipping cotton swabs in the 0.1% acriflavine solution and applying it to the oral mucosal location to be imaged.

The fluorescein solution in this study was prepared by diluting 1g/10mL vial of fluorescein sodium salt (Retinofluor 10% injection, Phebra Pty Ltd) with 10 ml of sterile water using a fine insulin syringe (0.5ml) using 0.1 ml of 1.0% fluorescein (1g/10ml) diluted in 10 ml of water. The fluorescein solution was provided to the imaging participant in the form of a 0.1% solution in a water cup which was topically administered by the patient swishing that solution.

Cellular architectural features were expected to be visualized with fluorescein, while nuclei were expected to be visualized with acriflavine.

3.2.4. Imaging protocol

The entire clinical imaging protocol of in vivo confocal microscope used in this study has been published as Yap et al. 2023 (Yap et al., 2023) (**Protocol 1**).

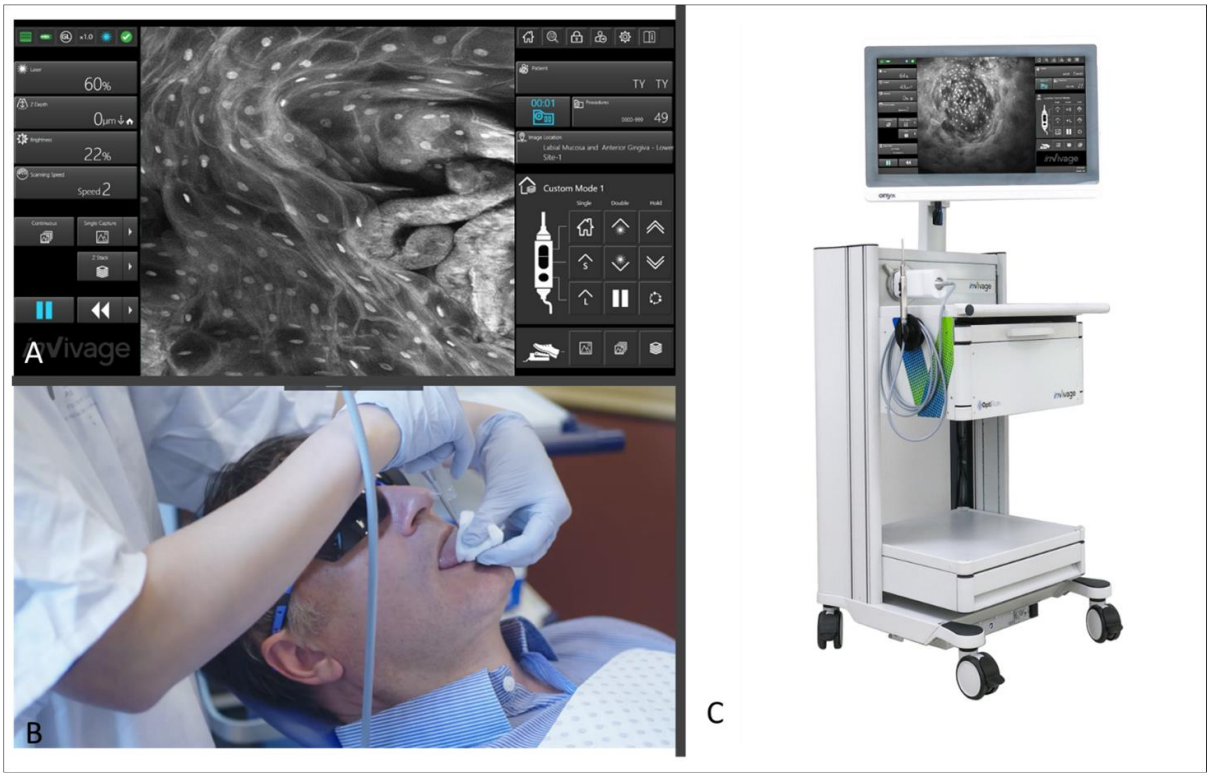


Figure 3.2: InVivage in vivo confocal microscopy image acquisition as adapted from Yap et al. (2023). (A) Software interface of the CLE, (B) Handheld probe chairside usage, (C) The portable InVivage unit

Protocol 1: In vivo clinical imaging protocol with confocal microscopy of the oral mucosa using topical contrast agents

1. Oral Cavity Examination

An oral medicine specialist performs a systematic inspection of all mucosal surfaces in the oral cavity.

2. Lesion Documentation

Register all intraoral lesions using macrographic photography.

Equipment: Canon EOS 200D with Canon 100mm F2.8 Macro USM lens

3. Confocal Microscope Imaging

Use the InVivage® confocal microscope for imaging. Patients were imaged with both of the following agents, in this order:

a) Fluorescein Imaging

1. Prepare a 0.1% solution of fluorescein in sterile water.
2. Instruct the patient to swish 10 mL of the solution in their mouth for 1 minute.
3. Ask the patient to rinse with water.
4. Perform confocal microscopy imaging.
5. Repeat the water rinse until the rinsed water appears clear.

b) Acriflavine Imaging

1. Prepare a 0.1% solution of acriflavine in sterile water.
2. Soak a cotton swab in the acriflavine solution.
3. Paint the mucosal areas of interest with the soaked cotton swab.
4. Instruct the patient to rinse with water for 1 minute.
5. Perform confocal microscopy imaging.
6. Repeat the water rinse until the rinsed water appears clear.

Image location-site selection was dependent on the intraoral location of mucosal abnormality. Image locations were defined as specific intraoral regions which, taken together, encompassed the entire oral cavity through macrographic photography, and were compatible with the mapping software developed concurrently called MouthMap™ (Figure 3.3.).

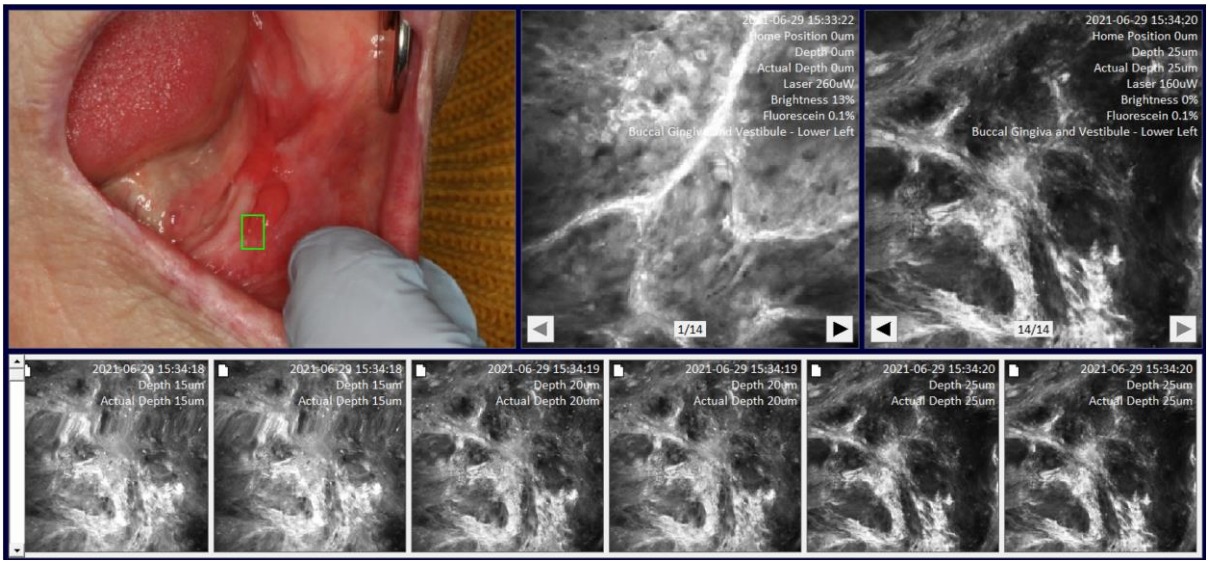


Figure 3.3. MouthMap™ visualisation for annotation of CLE images as adapted from (Yap et al., 2023).

Image site was defined as the specific point at which the probe was placed when the image capture occurred. An image set was defined as a set of images acquired using a single fluorescence agent collated by location and sites. Buccal mucosa or contralateral healthy tissue surfaces were used as control location-sites for each patient.

Capture at an image site was made by placing the probe on the site of interest and capturing a Z-stack. The Z-stack image acquisition commenced at the surface of the probe into the tissue until the depth of 400µm with 5 µm incremental focal plane advancements. This Z-stack command occurred utilising a programmed single foot pedal command so that the user could remain otherwise immobile whilst supporting the intraoral tissues. For all patient imaging, a long pass filter of 515– 815 nm was used, along with a scanning speed of 1024 pixels x 512 lines with a frame rate of 0.7 seconds per frame. For all the experiments undertaken in this study the term ‘image’ refers to a single 2D plane of data.

A standard-of-care scalpel biopsy and histopathological diagnosis were performed at the discretion of the attending clinician at the oral medicine clinic. Biopsies were conducted when histopathology was deemed necessary to confirm the diagnosis of a new or altered clinical presentation. The biopsy site was chosen according to standard protocols and prior to the use of the confocal microscope.

All subsequent analyses and machine learning experiments used random subsets of these contrast agent datasets for training, validation, and testing, ensuring consistency in the source population of images. All images presented

were processed and labelled according to community guidelines for microscopy image publication, ensuring appropriate scale representation and resolution in the final version (Schmied et al., 2024).

3.2.5. Diagnostic categories

The diagnostic triage models in this thesis were based on the detection of oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC). In an attempt to provide meaningful information to clinicians interpreting the diagnostic predictions of the models developed, 4 categories were established: 'no dysplasia', 'lichenoid', 'low-risk', and 'high-risk' (Figure 3.4., Table 3.2.).

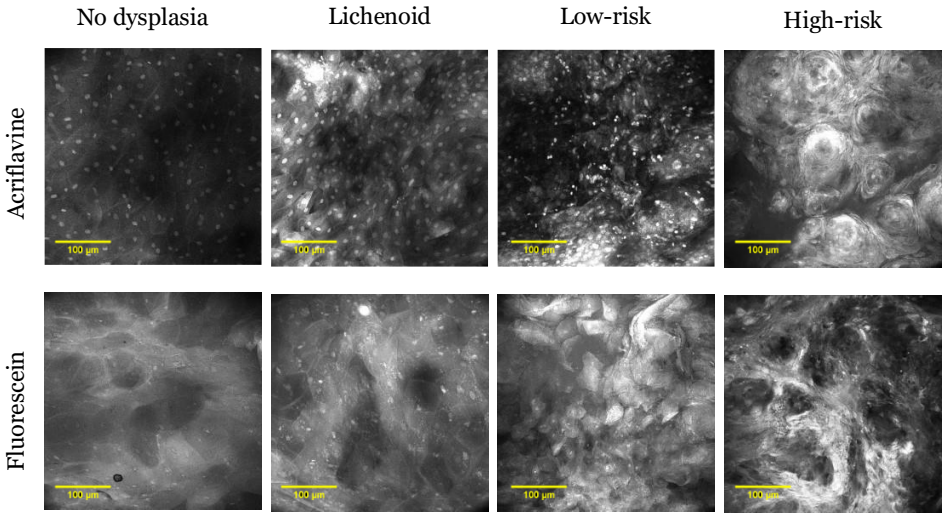


Figure 3.4. Examples of acriflavine and fluorescein images across the 4 disease categories

The 'No dysplasia' category contained lesions that were histopathologically diagnosed as normal oral epithelium in terms of dysplasia and oral cancer. Due to the similar presentation of oral lichen planus (OLP) and oral lichenoid lesions (OLL), clear distinction from oral epithelial dysplasia (OED), and their controversial malignant potential these lesions were considered as a separate 'lichenoid' category in this study (Gonzalez-Moles et al., 2008; Van der Meij & Van der Waal, 2003; Isaac Van der Waal, 2009). The remaining lesions were classified into 'low-risk' and 'high-risk' using the binary dysplasia grading for OED and OSCC as proposed by Kujan et al. (2006) and adopted by the WHO (Kujan et al., 2006). The binary system categorises lesions based on the overall number of observed cytological and architectural features listed in the WHO 2005 grading system (Kujan et al., 2006).

Table 3.2. Diagnostic categories for ML development based on histopathology

Diagnosis categories	Histopathology diagnosis
No dysplasia	Amalgam tattoo
	Chronic inflammation
	Denture associated gingival hyperplasia
	Fibroepithelial polyp
	Focal papillomatosis
	Hyperplasia & hyperkeratosis
	Squamous papilloma
	Verruciform xanthoma
	Histopathologically normal tissue
Lichenoid	Lichenoid inflammation
	Oral lichen planus
Low-risk	Atypia
	Low grade dysplasia
	Verrucous hyperplasia
High-risk	High grade dysplasia
	OSCC

3.2.6. Image analysis hardware

All machine learning experiments were conducted using a combination of local and cloud-based computing resources.

Local experiments were run on a desktop computer provided by the University of Melbourne. This device was equipped with an 12th Gen Intel Core i7-12700H, 2300 Mhz processor (Intel Corporation, USA), 16 GB DDR4 random access memory (RAM), and an NVIDIA GeForce RTX 3060 graphics processing unit (GPU) (NVIDIA Corporation, USA) running the Windows 10

Education operating system (Microsoft Corporation, USA). These experiments were primarily executed through the Spyder IDE, leveraging local computational resources.

For more computationally intensive tasks, particularly deep learning segmentation experiments involving CNNs, Google Colaboratory Pro (Google LLC, USA) was utilised, which provided access to an NVIDIA A100 GPU (NVIDIA Corporation, USA) and a pre-configured high-performance cloud environment. This combination allowed for efficient execution of both lightweight classical machine learning models and resource-intensive deep learning models. All classification speed measurements, particularly for CNNs, should be interpreted in the context of the respective hardware used during inference.

3.3. Study 1 – Micrograph Quality Filtering

This preliminary study was conducted for the purpose of filtering out the noise in the large image database while retaining the diagnostic quality images suitable for further analysis. All machine learning studies in this dissertation align with the STARD checklist for reporting diagnostic accuracy and the WHO-ITU checklist for artificial intelligence research in dentistry (Bossuyt et al., 2015; Schwendicke et al., 2021).

The CNNs developed in this study were trained and tested in two different environments for assessing their influence on model performance:

- MATLAB (MathWorks, USA)
- PyTorch (Python library)

MATLAB is often employed in educational and clinical research contexts where simplicity and rapid prototyping are prioritized, making its well-tuned default settings a practical baseline for typical users (Kim, 2017). In contrast, PyTorch is a flexible research framework that encourages fine-grained control and methodological rigor, justifying the use of hyperparameter optimization and cross-validation (Paszke et al., 2019). Practical constraints such as computational efficiency and ease of experimentation differentiate these development environments. MATLAB lacks streamlined tools for large-scale grid search and cross-validation, whereas PyTorch integrates seamlessly with such workflows. This design of developing CNN models in two different environments enables a comparative analysis that reveals how much performance improvement is achievable through modern tuning techniques.

All programming code developed as a part of all studies described in this dissertation are available on the GitHub repository created for this work and is addressed in Appendix 2.

Objectives:

1. To develop, train, validate, and test a convolutional neural network model that can identify diagnostic quality, in vivo fluorescence confocal microscopy images in the MATLAB deep learning development environment.
2. To develop, train, validate, and test a convolutional neural network model that can identify diagnostic quality, in vivo fluorescence confocal microscopy images in the PyTorch deep learning development environment using hyperparameter optimisation and cross validation.

3. To use this diagnostic quality filtering pipeline on the entire acriflavine and fluorescein database to create a dataset consisting solely of images of diagnostic quality.

The quality filtering CNNs termed as Quality Micrograph Refiners (QMR) were designed to filter out confocal micrograph images that were of poor quality. The criteria used for including images of sufficient quality were (Figure 3.5.):

1. Presence of visible and in focus oral epithelial cell borders or oral epithelial cell nuclei (Figure 3.5.b).
2. Absence of major artifacts that cover equal to or more than 75% of the field of view of the confocal micrograph (Figure 3.5.c).
3. Absence of major imaging errors that cover equal to or more than 75% of the field of view of the confocal micrograph (Figure 3.5.d).
4. Absence of featureless zones that cover equal to or more than 75% of the field of view of the confocal micrograph (Figure 3.5.e).

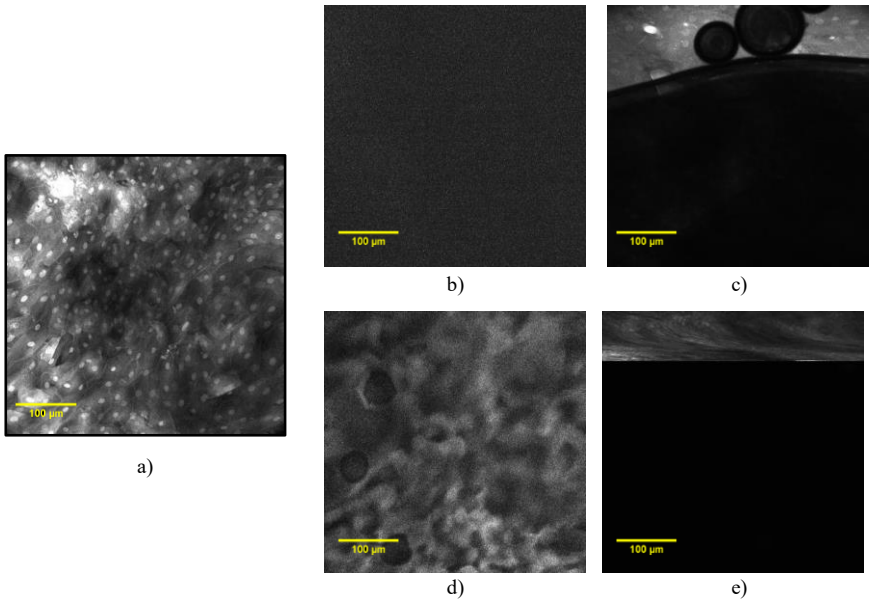


Figure 3.5. Examples of confocal micrographs used for quality filtering CNN development based on inclusion criteria; a) Example of diagnostic quality image that fulfils all criteria, b) Absence of cell borders or nuclei, c) Major artifact (water/saliva bubble), d) Major imaging error (out of focus), e) Major featureless zone

3.3.1. Quality filtering CNN development in MATLAB

A convolutional neural network (CNN) architecture was developed by including CLE images as the input and classifying them as being either of diagnostic quality or being of poor quality. The quality filtering CNN developed in this experiment is named the 'MATLAB QMR' where QMR stands for Quality Micrograph Refiner and it was developed in the MATLAB (MathWorks, USA) software's Deep Network Designer application.

3.3.1.1. Development framework

This model was developed within MATLAB's (MathWorks, USA) Deep Network Designer module. The Inception_v3 architecture developed by Szegedy et al. (2015) and pre-trained on the ImageNet dataset was used as the CNN architecture (Szegedy et al., 2016) (Figure 3.6.).

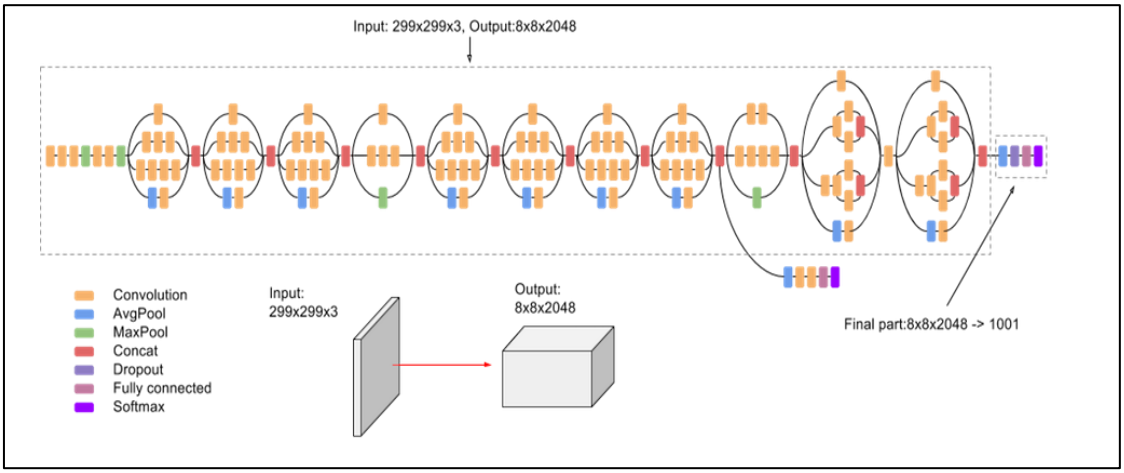


Figure 3.6. Inception_V3 CNN architecture as designed by and adapted from Szegedy et al. (2015)

3.3.1.2. Transfer learning

Transfer learning was used to develop this MATLAB QMR model from the pre-trained Inception_V3 by repurposing its classification target to confocal microscopy images. For transfer learning two changes were made to the classification head of the model architecture (Figure 3.7.).

The last fully connected layer was replaced with a new fully connected layer that has the output classes required for the current study. The final classification output layer was replaced with a new one that used the cross-entropy loss function to assist in classifying images into the defined classes (instead of the original 1000).

3.3.1.3. Training parameters

The training parameters selected for this model were the default values recommended by the Deep Network Designer module. The number of epochs that signifies how many times the model iterates over all the training images was set to the default value of 30. The learning rate that specifies the rate at which the model updates its internal parameters was set at the default value of 0.01.

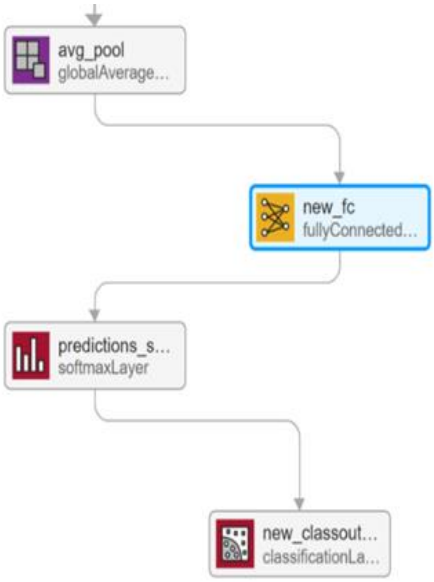


Figure 3.7. Replacing the last fully connected layer in the Inception_V3 architecture with a new custom layer labelled 'new_fc' which repurposes the model to the task of quality filtering within MATLAB's Deep Network Designer tool

The default solving algorithm for the training process was the stochastic gradient descent (SGD), with the momentum set to '0.9'. The SGD is a commonly used optimization algorithm in deep learning that minimizes the loss function computed by the model by iteratively adjusting the model's internal parameters (LeCun et al., 1998). Instead of looking at the entire training dataset at once to update the model's parameters, SGD randomly selects a small subset of the dataset, called a minibatch. In each iteration, the gradient of the loss function is calculated based on one minibatch, and the model's parameters are updated accordingly. The gradient estimates how the model's internal parameters need to be changed to reduce the loss.

The parameters are then updated in the direction that reduces the cross entropy loss function, according to the calculated gradient and a learning rate, which controls the size of the update step (LeCun, Bengio, & Hinton, 2015b). Once all minibatches have been used for updates, one epoch is complete. Using minibatches lowers the computational cost by processing fewer data

points per iteration with the benefit of smoother and more stable convergence towards low loss. Thus, it is faster and more memory-efficient than using the full dataset (LeCun et al., 2015b).

3.3.1.4. Pre-processing

Prior to any analysis using the machine learning techniques described across all the studies in this dissertation the images were pre-processed. The pre-processing steps were carried out using the open-source image analysis software Fiji (ImageJ) (Schindelin et al., 2012). These steps are described within Protocol 2.

Protocol 2: Steps for pre-processing images before using Fiji, ImageJ software

1. Conversion from compressed DICOM to uncompressed TIFF images required for image classification.

2. Conversion of TIFF images from greyscale (1 channel) to RGB (3 channels) that is an input image requirement of Inception_V3.

3. Conversion of image size from 1024x1024x3 to 299x299x3 using bicubic interpolation.

3.3.1.5. Dataset curation

For development of the MATLAB QMR model, 800 randomly selected images with 400 taken from each contrast agent dataset representing the intra-oral tissue of buccal mucosa, tongue, remaining keratinised mucosa and remaining non-keratinised mucosa were selected.

All of these images were manually annotated as ‘unusable’ and ‘usable’ by blind screening and then consensus decision between three investigators. A final 531 images were annotated ‘unusable’ and 269 were annotated ‘usable’. Among the 800 images chosen for the training phase 80% were used for learning (n=640) and 20% were internal validation images (n=160).

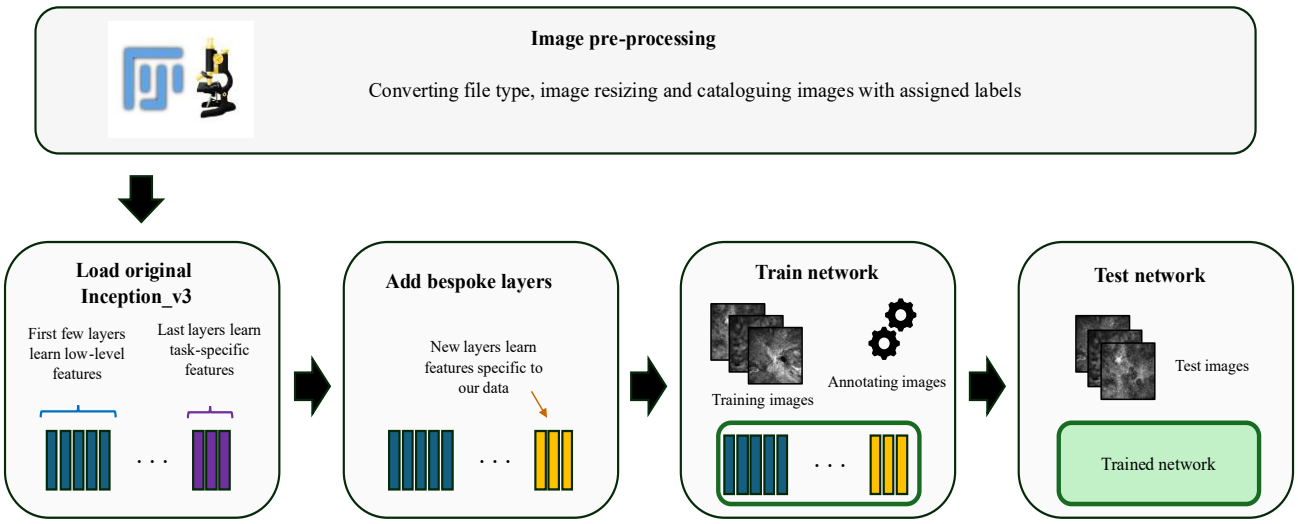


Figure 3.8. CNN development workflow using transfer learning on the Inception_V3 architecture

The images were pre-processed (Protocol 2), annotated, the inception_V3 model was modified, training was carried out and then the model was tested on unseen images (Figure 3.8.). The bespoke modified Inception_V3 model was trained for 960 iterations over 30 epochs in mini batches of 20 images (32 iterations per epoch). This training process was validated every 10 iterations against the 160 internal validation images.

3.3.1.6. Performance assessment

To test the MATLAB QMR a test dataset was constructed with 400 previously unseen images comprising 50 images from each of the 8 site representative groups mentioned above. These test images had been manually evaluated and annotated by blind screening and then consensus decision between three investigators (one general dentist and 2 oral medicine specialists). The trained MATLAB QMR model was evaluated based on its classification performance on the test dataset using a confusion matrix. Performance metrics of accuracy, sensitivity, specificity, precision, and F1 score were calculated (Table 3.3).

Table 3.3. Performance metrics for assessing machine learning model performance

Metric	**Definition**	**Formula**
Accuracy	The proportion of correct predictions (both true positives and true negatives) out of the total predictions.	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity/Recall	The proportion of actual positives that the model correctly identifies.	$Sensitivity = \frac{TP}{TP + FN}$
Specificity	The proportion of actual negatives that the model correctly identifies.	$Specificity = \frac{TN}{TN + FP}$
Precision	The proportion of predicted positives that are actual positives.	$Precision = \frac{TP}{TP + FP}$
F1 score	The harmonic mean of precision and recall, balancing the two.	$F1\ score = 2 * \frac{Precision\ .\ Recall}{Precision + Recall}$
Receiver operator characteristic (ROC)	A plot of the true positive rate (TPR) against the false positive rate (FPR) across different thresholds.	$TPR = \frac{TP}{TP + FN}$ $FPR = \frac{FP}{FP + TN}$
Area under the ROC curve (AUROC)	A single value summarizing the ROC curve, indicating the model's ability to distinguish between classes	$AUROC = \int_0^1 TPR(FPR)d(FPR)$
TP = True positives, TN = True negatives, FP = False positives, FN = False negatives		

The test performance metrics were calculated for all acriflavine images, all fluorescein images, and images from the buccal mucosa, floor of mouth, gingiva & vestibule, hard palate, soft palate, and tongue individually.

3.3.2. Quality filtering CNN development in PyTorch

This experiment involved developing a quality filtering CNN named the 'PyTorch QMR' where QMR stands for Quality Micrograph Refiner. This experiment used the PyTorch deep learning framework in the Python programming language and utilised certain libraries and packages to provide a robust hyperparameter optimised and cross validated development process.

Several different variations of the quality filtering CNN models were developed, ranked, with the developed best being selected to represent PyTorch QMR.

3.3.2.1. Development framework

The CNNs in this experiment were developed within the PyTorch (ver. 2.1.0) framework using Python 3 programming language (ver. 3.11.5) (Paszke et al., 2019; Guido van Rossum & FL Drake, 2021). Python libraries of Numpy (ver. 1.24.3) and Pandas (ver. 2.1.1.) were used for performing numerical calculations and the Sci-kit learn library (ver. 1.3.0.) was used for machine learning applications (McKinney, 2011; Pedregosa et al., 2011a; Van Der Walt, Colbert, & Varoquaux, 2011). The CNN architecture in this study was a modified Inception_V3 architecture where the final fully connected layer was replaced by a new fully connected layer for the classes required in this study using transfer learning.

Before the images could be loaded into a PyTorch environment they underwent pre-processing steps using Protocol 2. The first step in the training and evaluation process was converting the data into PyTorch tensors that are a fundamental multidimensional data structure in PyTorch designed for efficient numerical computation (Paszke et al., 2019). The Inception_V3 model has a specific input image requirement of 299 x 299-pixel RGB images. A custom dataset Python class was defined in this study to pre-process and resize the images along with the 'torchvision.transforms' provided by PyTorch. The learning algorithm used was stochastic gradient descent (SGD) with momentum of 0.9.

3.3.2.2. Hyperparameter optimisation

A grid search approach was employed for hyperparameter optimization across a spectrum of number of epochs and learning rate. This method involved creating a table of hyperparameter values and systematically searching through all possible combinations of these values to determine the combination that results in the best performance of the model. Values of '5', '10' and '15' epochs and learning rates of '0.001', '0.01', '0.1', were explored (Bergstra & Bengio, 2012). The training dataset was split into training and validation sets using K fold cross validation where K=5. This resulted in each cross-validation fold having 80% of the data constituting the training set and 20% the validation set with the data being shuffled across all folds. One model was trained for each combination of epoch and learning rate for each cross-validation fold. This resulted in a total of 45 CNN models being trained and

tested across all combinations of epochs, learning rates and cross validation folds. The MATLAB CNN was trained using the default hyperparameter values of 30 epochs and learning rate of 0.001.

3.3.2.3. Performance assessment

The performance metrics used to assess the CNN models in this study were accuracy, sensitivity, specificity, precision, and F1 score (Table 3.2.). The accuracy results across all folds of the 5-fold cross-validation were averaged to estimate the most optimum hyperparameter combination. All trained models were ranked based on an aggregation of ranks for 5 metric scores calculated using Protocol 3.

Protocol 3: Ranking of ML models based on their test performance to determine the best model

1) Performance Categorization

Categorize the performance of all trained models separately for each class.

2) Cross-Validation Ranking

a) For models from individual folds of cross-validation:

   i) Calculate overall aggregate ranks across all 5 metrics.

   ii) Calculate ranks for all classes (one vs all).

b) Identify the rank #1 model, which represents the best performing model for predicting all classes.

3) Overall Rank Calculation

a) Rank the aggregate rank scores for each hyperparameter combination.

b) Use these rankings to determine the overall rank of each model.

4) ROC Analysis

For each diagnostic class vs all:

a) Plot the receiver operator characteristic (ROC) curve.

b) Calculate the area under the ROC curve (AUC).

The 5 folds were numbered from 0 to 4 as python indexing begins from 0. Performance metrics for the ML model in terms of intraoral location and

disease category were calculated individually. All statistical analyses were performed using the Python sci-kit learn library (Pedregosa et al., 2011a).

Further assessment of the best ranking model during testing was undertaken by assessing different subsets of the test dataset based on contrast agent and the 6 intraoral sites that represent keratinised and non-keratinised tissue. Therefore, test performance metrics were calculated for all acriflavine images, all fluorescein images, and images from the buccal mucosa, floor of mouth, gingiva & vestibule, hard palate, soft palate, and tongue, individually. The performance of this model was compared to that of the quality filtering model developed in MATLAB QMR.

The best performing model was selected as the PyTorch QMR model (Code structure outline 1).

Code structure outline 1: PyTorch python script for training CNN models using hyperparameter optimisation and cross validation

Import necessary libraries and define paths:

a) Import libraries for image processing, deep learning, metrics, and file handling.

b) Define file paths for saving models and data directories.

2) Define a custom dataset class:

a) Initialize dataset with root directory and transformations.

b) Map class names to labels and load data by iterating over class directories.

c) Implement `__len__` for dataset size and `__getitem__` for retrieving data samples.

3) Define a training and evaluation function:

a) Set up model on GPU if available.

b) Initialize variables for tracking training/validation losses and accuracy.

c) Iterate through epochs:

d) Training phase:

i) Set model to training mode.

ii) Loop through batches, perform forward pass, compute loss, backpropagate, and update weights.

iii) Track training loss and accuracy.

e) Validation phase:

i) Set model to evaluation mode.

ii) Loop through validation batches, perform forward pass, compute loss and predictions.

iii) Track validation loss and accuracy.

f) Save the model with the best validation accuracy.

4) Preprocess images:

a) Define image resizing, normalization, and augmentation transformations.

5) Initialize the dataset and data loader:

a) Create a custom dataset instance and split data into train/validation sets using `StratifiedKFold`.

6) Define hyperparameters:

a) Set ranges for epochs and learning rates.

7) Perform training for different hyperparameter combinations:

a) Iterate over folds from `StratifiedKFold`.

b) For each fold, create train and validation datasets.

c) Iterate over epochs and learning rates:

i) Load a pre-trained model (InceptionV3) and modify for transfer learning.

ii) Define optimizer and loss function.

iii) Train and evaluate the model.

iv) Save the trained model and its metrics.

v) Compute confusion matrix and derived metrics (e.g., accuracy, precision, F1 score).

vi) Store results in a dataframe.

8) Save results:

a) Export results to an Excel file.

9) Calculate and print elapsed time for the entire process.

3.4. Study 2 – Machine learning analysis of human identified qualitative features

This study involved the use of human observed and annotated qualitative features relating to cells, nuclei and fluorescent granules in confocal microscopy images with machine learning models for the diagnostic triage of oral epithelial dysplasia and oral squamous cell carcinoma.

Objective:

1. To train, validate and test different machine learning models for identifying oral potentially malignant disorders in the form of lichenoid lesions, oral epithelial dysplasia, along with oral squamous cell carcinoma based on qualitative human identified features observed on in vivo captured fluorescence confocal endomicroscopy images.

This study included a data subset of 600 images (n=300 per contrast agent) captured in 51 patients from the buccal mucosa, hard palate, soft palate, tongue. The 4 diagnostic categories for ML development in this study were assigned based on the histopathology diagnosis of each imaged lesion (Table 3.2.).

3.4.1. Qualitative features selection

The acriflavine and fluorescein images were screened to determine the presence and absence of identifiable cellular and nuclear features by two investigators. The assessment of these features was undertaken using a categorical scoring system. Features of epithelial cells, nuclei and cytoplasm were included in the scoring (Table 3.4). The images were pre-processed for analysis using Protocol 2 in Section 3.3.

These qualitative feature variables needed to be converted to fixed length categorical variables for use as inputs into ML models for diagnostic prediction. These features were scored in 2 different tiered systems for generating two feature sets. The evaluation of these features was conducted using two distinct systems: a multi-tier system and a binary system. In the multi-tier system, each feature was described using one of several possible categories based on its appearance in the analysed image. In contrast, the binary system assigned one of two possible values (i.e. 0 & 1) to each feature, indicating its presence or absence in the image (example in Figure 3.9.). A score of ‘1’ indicated the presence of said features in that image and ‘0’ denoting its absence.

All multi-tier observations were grouped in ‘Feature set A’ and the binary observations for these features were grouped in ‘Feature set B’ (Table 3.4., Figure 3.9.).

Table 3.4. Qualitative features assessed for ML diagnostic prediction analysis

Feature name	**Feature set A variables (multi)**	**Feature set B variables (binary)**
Cell crowding	0 = Absent, 1 = Low, 2 = High	0 = Absent, 1 = Present
Cell size homogeneity	0 = Regular, 1 = Borderline irregular, 2= Mildly irregular, 3 = Moderately irregular, 4 = Highly irregular	0 = Regular, 1 = Irregular
Nuclei crowding	0 = Absent, 1 = Low, 2 = High	0 = Absent, 1 = Present
Nuclei size homogeneity	0 = Regular, 1 = Borderline irregular, 2= Mildly irregular, 3 = Moderately irregular, 4 = Highly irregular	0 = Regular, 1 = Irregular
Fluorescent granules	0 = Absent, 1 = Limited presence, 2= Abundantly present	0 = Absent, 1 = Present

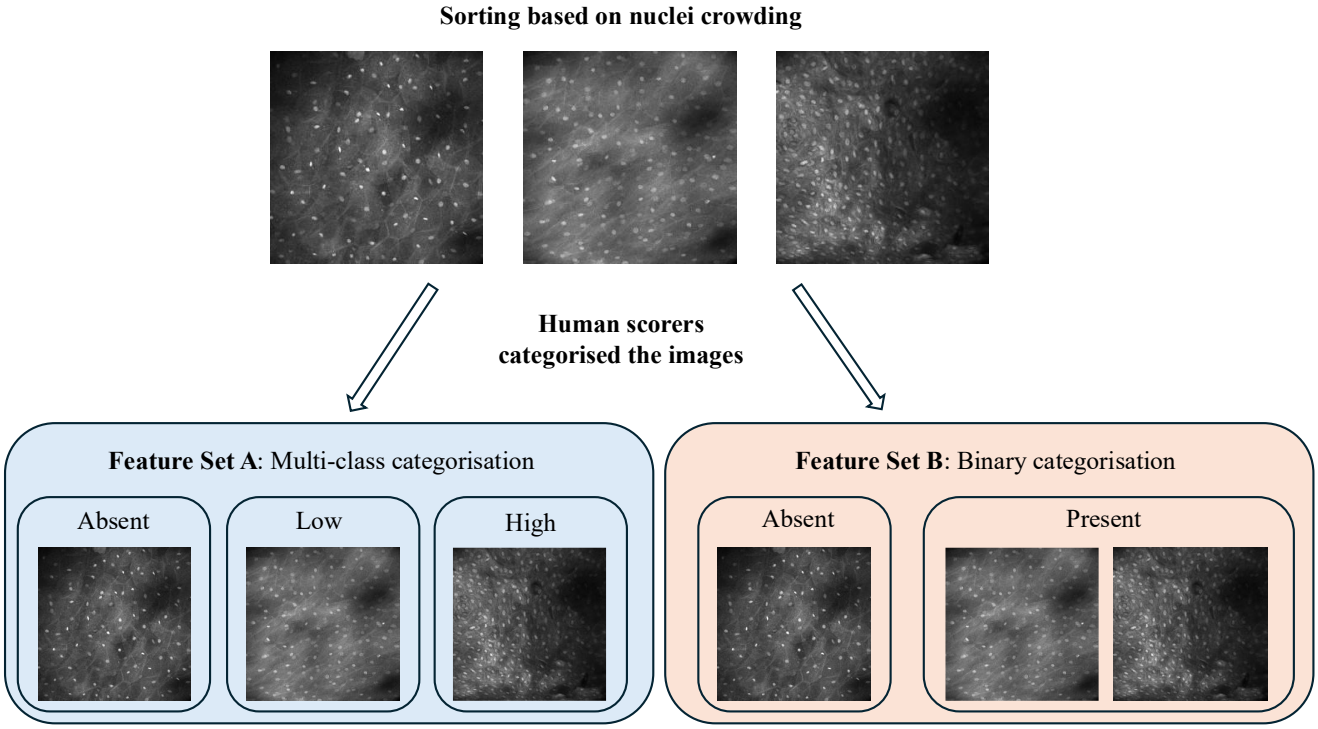


Figure 3.9. Example of sorting confocal micrographs by human scorers based on the observed nuclei crowding into multiple classes for Feature set A and binary classes for Feature set B

These specific features were selected based on histopathology features of OED and OSCC and on the confocal microscopy features observed in image frames of the oral mucosal epithelium (Ramani et al., 2023; Reibel et al., 2017b; Yap et al., 2023). These 5 selected features represent epithelial cell, nucleus, and cytoplasm characteristics while maintaining relatively low feature dimensionality.

Fewer features prevent the ML models from becoming overly complex that could lead to model's overfitting on the training data and showing poor generalisation on evaluation.

3.4.2. Chi-squared feature analysis

A chi-squared analysis was carried out to determine whether there was a statistically significant association between the feature variables and the categorical target variable. The chi-squared test was selected to provide an evaluation on how strongly each observed feature was related to the diagnostic category classes.

To prepare the dataset for the chi-squared tests all feature data from both feature sets for both acriflavine and fluorescein datasets were binned into categories. The limitation of this test of not being able to indicate the direction or nature of any relationships was acknowledged.

The outputs of the chi-squared tests such as the chi-squared statistic and p-value were used to interpret the discrepancies between the observed data and expected data assuming the data is unrelated to the classes in order to assess the presence of relationships between each feature and the diagnostic classes.

3.4.3. Machine learning development

Four machine learning approaches of logistic regression, support vector machines (SVM), random forest (RF), and XGBoost (XGB) were used to classify qualitative image feature data samples into the four classes: ‘no dysplasia’, ‘lichenoid’, ‘low-risk’, and ‘high-risk’ (Table 3.2.). The aim was to evaluate and compare the performance of these models in accurately predicting the diagnostic classes.

3.4.3.1. Model Selection

Four machine learning models were selected for this study, each offering unique advantages in classification tasks (Table 3.5.).

1. Logistic Regression (LR): A simple, interpretable model often used as a baseline for classification (Kleinbaum, Dietz, Gail, Klein, & Klein, 2002).
2. Support Vector Machine (SVM): A model that constructs hyperplanes to separate classes in a high-dimensional space, effective for non-linear separations when paired with kernel functions (Noble, 2006).
3. Random Forest (RF): An ensemble model that combines multiple decision trees to improve prediction accuracy and reduce overfitting (Breiman, 2001).
4. XGBoost (XGB): An ensemble model known for its high efficiency and performance, particularly in structured data tasks (Chen & Guestrin, 2016).

Table 3.5. Machine learning models comparison of properties

	Logistic Regression	**Support Vector Machine (SVM)**	**Random Forest (RF)**	**XGBoost (XGB)**
Type	Linear model	Non-linear model	Ensemble (bagging) model	Ensemble (boosting) model
Concept	Predicts probabilities using a linear equation	Maximizes margin between classes with hyperplanes	Builds multiple decision trees and averages them	Sequentially builds trees to reduce errors
Complexity	Low	Medium to High	Medium	High
Data Handling	Works best with linearly separable data	Works well with non-linear data	Handles both linear and non-linear data	Excels with complex and large datasets
Feature Importance	Coefficients represent feature importance	Implicit; not easily interpretable	Provides feature importanc e	Provides feature importance
Overfitting Risk	Low (with regularizatio n)	Medium (if not tuned properly)	Low (due to averaging)	Low (due to regularizatio n and boosting)
Training Speed	Fast	Slow for large datasets	Moderate	Slower than Random Forest
Prediction Speed	Fast	Moderate	Fast	Moderate

Hyperparameter Tuning	Few parameters to tune	Needs careful tuning (kernel, C, gamma, etc.)	Moderate tuning required	Many parameters to tune
Scalability	Highly scalable for large datasets	Not very scalable	Scalable with parallel processing	Highly scalable with distributed computing
Interpretability	High	Low	Moderate	Low
Use Cases	Binary classification, e.g., fraud detection	Image classification, text categorization	Complex problems like healthcare	Highly competitive scenarios like competitions

3.4.3.2. Hyperparameter Optimization

To enhance model performance, hyperparameter optimization was performed using grid search. This approach involved systematically testing combinations of predefined hyperparameters and selecting the configuration that maximized cross-validated performance. Key hyperparameters tuned for each model included:

1. **LR**: Regularization strength C (0.1,1,10) and penalty
2. **SVM**: Kernel type, regularization parameter (C), and kernel-specific parameters (e.g., gamma for RBF kernel).
3. **RF**: Number of trees, maximum depth, and minimum samples required for splits.
4. **XGB**: Learning rate, number of estimators, maximum depth, and subsampling rate.

3.4.3.3. Model Training and Evaluation

Each model was trained using the training subset and evaluated on the testing subset. The following evaluation metrics were calculated to assess model performance (Table 3.3.):

1. **Accuracy**: The proportion of correct predictions out of total predictions.
2. **Sensitivity (Recall)**: The ability of the model to correctly identify true positives for each class.
3. **Specificity**: The ability of the model to correctly identify true negatives for each class.
4. **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.
5. **F1 Score**: The harmonic mean of precision and recall, offering a balanced measure of performance.

The test performance results of all models were evaluated using Protocol 3 to determine the best performing model (rank #1).

3.4.3.4. Software and Tools

All analyses were conducted in Python using the following libraries:

1. **scikit-learn:** For implementing LR, SVM, and RF models, as well as preprocessing and evaluation metrics.
2. **XGBoost:** For training and tuning the XGB model.
3. **pandas and numpy**: For data manipulation and preprocessing.

This methodological approach ensured a systematic comparison of the models under consistent conditions, providing robust insights into their suitability for multi-class classification tasks involving categorical data (Code structure outline 2).

Code structure outline 2: Development of the logistic regression, SVM, random forest, and XGBoost models using hyperparameter optimisation and cross validation on the data from excel worksheets

1. Import Necessary Libraries
    a. ‘pandas’, ‘numpy’ for data manipulation.
    b. ‘scikit-learn’ for machine learning models, preprocessing, and evaluation.
    c. ‘xgboost’ for the XGBoost model.
2. Load Excel Workbook

    a. Read the Excel workbook using pandas.read_excel().
    b. Load training data and test data from their respective sheets.
3. Extract Features and Class Labels
    a. Identify feature columns and the class label column.
    b. Separate features (X) and target class values (y) for both training and test datasets.
4. Z-Score Standardisation
    a. Use StandardScaler from sklearn.preprocessing to standardize the feature values.
5. Train-Test Split for Cross-Validation
    a. Combine training and test data for splitting using an 80:20 ratio.
    b. Use train_test_split from sklearn.model_selection for splitting.
6. Set Up Machine Learning Models
    a. Define Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost models.
    b. Specify the parameter grid for each model for hyperparameter tuning.
7. Grid Search with Cross-Validation
    a. Use GridSearchCV with 3-fold cross-validation for hyperparameter tuning.
    b. Evaluate models using accuracy or another metric.
8. Fit and Test Models
    a. Train models on the training dataset.
    b. Test models on the testing dataset and evaluate performance.
9. Output Results
    a. Output the best hyperparameters and model performance metrics for each model.

3.5. Study 3 – Quantitative ML feature extraction analysis

This study focuses on extraction and measuring of quantitative features visualised in the fluorescence in vivo captured human confocal microscopy images to be used as inputs for ML models that would provide diagnostic classification of images into the diagnostic categories of 'no dysplasia', 'lichenoid', 'low-risk', and 'high-risk' (Table 3.2.).

The feature selected for extraction and measurement was epithelial cell nuclei. Nuclei are known to manifest changes in appearance and staining characteristics in histopathology for disorders, such as oral lichen planus (OLP), oral lichenoid lesions (OLL), OED, and OSCC (Warnakulasuriya, Kujan, Aguirre-Urizar, Bagan, González-Moles, Kerr, Lodi, Mello, Monteiro, Ogden, et al., 2021). Multiple measurements were extracted for cell nuclei for ML analysis.

Multiple experiments and approaches were explored in this study including preparation of images for feature extraction, training the feature extraction algorithm, curation of measurement variables, and developing ML models to use these measurements for diagnostic classification. The method of feature extraction described here is segmentation which involves the separation of regions of interest (ROI) from the background in an image.

3.5.1. Experiment 1 – Comparison of Cellpose 2D and StarDist 2D for optimal epithelial cell nuclei segmentation performance using a custom annotation process

This experiment involved creating a custom annotation process for cell nuclei in order to prepare the training and test datasets for all the ML feature extraction analysis downstream.

Two open source software packages compared in this experiment were Cellpose 2D and StarDist 2D (Schmidt, Weigert, Broaddus, & Myers, 2018; Stringer et al., 2021). Cellpose 2D is a generalist model for cellular segmentation that involves a process for annotation of ROIs within the software window (Stringer et al., 2021). StarDist 2D is a detection and segmentation algorithm for cells and nuclei in microscopy images based on the

identification of star-convex polygons as shape representation (Schmidt et al., 2018).

Objectives:

1. To develop a custom epithelial cell nuclei annotation process tailored for the feature extraction segmentation task.
2. To test and compare Cellpose 2D and StarDist 2D nucleus segmentation performance on a small fluorescence in vivo confocal microscopy dataset.

3.5.1.1. Custom annotation process

A total of 40 in vivo confocal microscopy images containing 10 images representing each of the disease classes: no dysplasia, lichenoid, low-risk and high-risk were collected for this initial experiment. The images were pre-processed for analysis using Protocol 2 in Section 3.3.

A new annotation pathway was trialled in this study. The images were initially annotated using the Cellpose software with the cyto3 generalist algorithm (pre-trained) to identify nuclei (Stringer & Pachitariu, 2025). This automated annotation process was complemented by a manual annotation step adding approximately 20-30 additional annotations per image and corrected any errors in the initial automated annotations (Figure 3.10.) (Stringer & Pachitariu, 2025). A total of 1820 nuclei were identified across all 40 images with 1273 nuclei in the training set and 547 nuclei in the test set.

These regions of interest (ROIs) were exported from the Cellpose software as .zip files. Subsequently, the images and their respective ROIs were imported into Fiji (ImageJ). Using the LOCI ROI mask generator, binary mask 8-bit TIFF files were created, depicting white nuclei against a black background.

For the purposes of model training and testing, 32 images (80%) were designated for training, while the remaining 8 images (20%) were allocated for testing. A power calculation was undertaken for the test image sample size for an effect size of 0.5 with an alpha of 0.05.

These images were uploaded to a secure private Google Drive for processing in Google Colaboratory notebooks using the following data structure:

1) **Training dataset**
    a) Images (Training source)
        i) img_1.tif, img_2.tif, ...

b) Label images (Training target)

i) img_1.tif, img_2.tif, ...

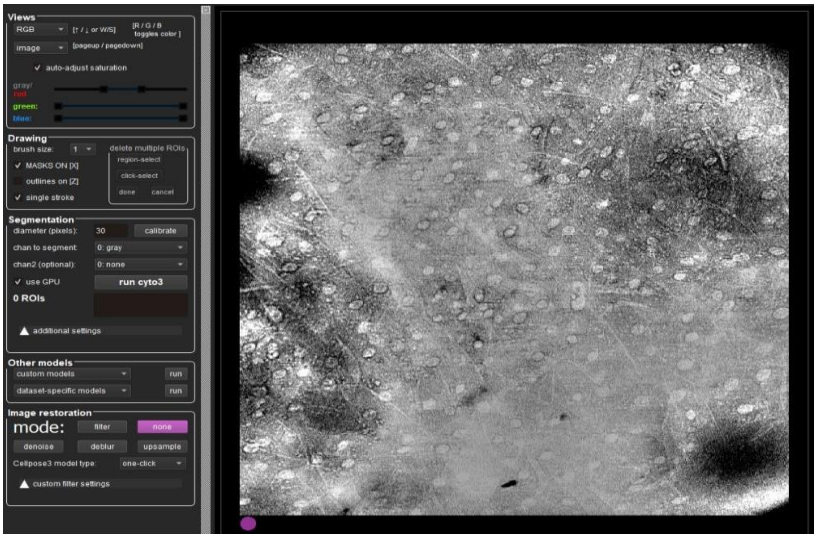
2) **Testing dataset**

a) Images

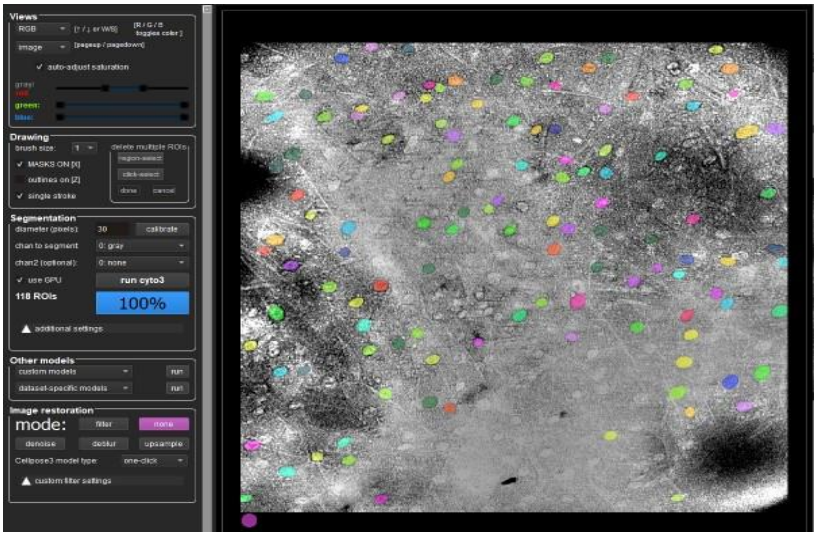
i) img_1.tif, img_2.tif

b) Masks

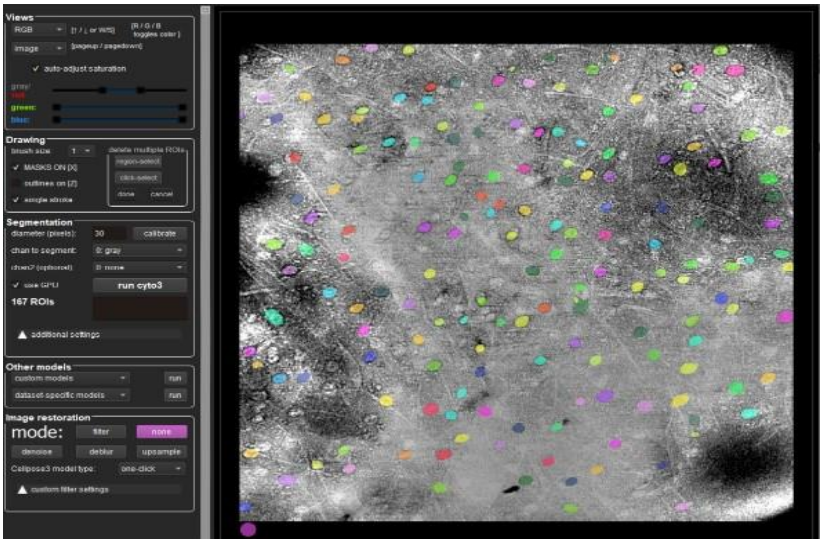
i) img_1.tif, img_2.tif



Step 1: Load image into Cellpose 2D



Step 2: Run default Cellpose Cyto3 segmentation model



Step 3: Human makes corrections and additions

Figure 3.10. Steps for custom annotation process for cell nuclei in Cellpose 2D

3.5.1.2. Segmentation model comparison

The training and validation was undertaken on the Google Colaboratory platform using Jupyter notebooks built by Henriques Lab at ZeroCostDL4Mic (von Chamier et al., 2021). These notebooks were specifically designed in the python programming language to train and validate Cellpose 2D and StarDist 2D models.

The ZeroCostDL4Mic notebook including all the programming code used for this experiment can be accessed online for free (von Chamier et al., 2021).

Code structure outline 3: ZeroCostDL4Mic notebooks for Cellpose 2D and StarDist 2D in Google Colaboratory

1. The notebook initiates by installing the Cellpose 2D/StarDist 2D libraries.
2. It then loads the necessary Python libraries required by Cellpose 2D/StarDist 2D. These libraries include numpy, matplotlib, pandas, csv, scipy, skimage, sklearn, and tqdm.
3. The notebook connects to a remote Graphics Processing Unit (GPU) online to accelerate training. In this experiment, an Nvidia Tesla T4 GPU was used, featuring 320 tensor cores and 16 GB of memory, which provides up to 40 times the processing power of a conventional CPU (nvidia white paper reference).
4. Users are able to connect to a Google Drive containing the training and test datasets.
5. The notebook allowed for the establishment of training parameters, such as the number of epochs, batch size, learning rate, and validation percentage.
6. The training process is executed, and the model weights are saved based on the best performance observed during training.
7. The trained model undergoes quality testing on an unseen test dataset.
8. The performance of the model is assessed using the following metrics: intersection over union (IoU), precision, and F1 score.

Table 3.6. Parameters used for training the Cellpose 2D and StarDist 2D models

Parameter	**Explanation**
Number of epochs	Number of complete passes made by the model through all the training images.
Batch size	The number of training samples seen by the model before it updates all its internal parameters.
Percentage validation	Proportion of images used to validate the model each epoch during training.
Initial learning rate	A hyperparameter used as a multiplier by the stochastic gradient descent optimizer during training.

The intersection over union (IoU) scores, precision, and F1 scores of both models were compared using mean and standard deviation. IoU is a widely used metric in object segmentation that is calculated as the ratio of the intersection area between the predicted and ground truth boxes to their union area. IoU scores range from 0 to 1, with 1 indicating perfect overlap and 0 indicating no overlap (Figure 3.11.).

Statistical analysis to compare performance of the Cellpose 2D and StarDist 2D models was undertaken using the paired sample t-test.

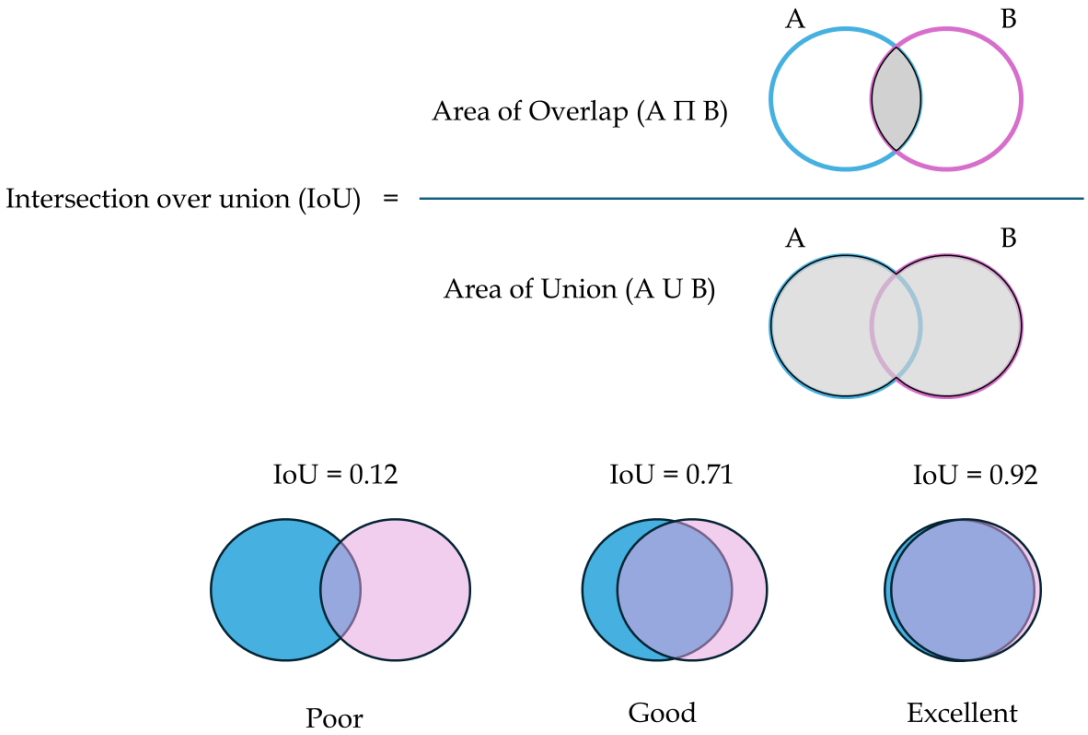


Figure 3.11. Calculation of the intersection over union (IoU) metric for object segmentation with examples of poor, good, and excellent IoU

3.5.2. Experiment 2 - Developing epithelial cell nuclei segmentation model for accurate feature extraction

This experiment involved the development of an nuclei segmentation algorithm that could be used to segment nuclei on fluorescence CLE images for analysis and diagnostic prediction of OPMDS in the form of lichenoid, and OED lesions along with OSCC lesions.

Objectives:

1. Design and implement a segmentation model tailored to fluorescence human in vivo captured confocal endomicroscopy images of the oral mucosa.
2. To evaluate segmentation model performance using metrics such as accuracy, F1 score, Mean true intersection over union score, Mean matched intersection over union score, and panoptic quality.

3.5.2.1. Dataset preparation using custom annotation process

An image data subset of 400 acriflavine and fluorescein in vivo confocal microscopy images from 14 participants were selected for training and testing a new nucleus segmentation model. The dataset was divided into 360 training (90%) and 40 testing images (10%).

The annotation of these images was carried out on a modified human-in-the-loop approach in Cellpose 2D (Pachitariu & Stringer, 2022). This involved using the default 'Cyto 3' model in Cellpose 2D followed by human intervention in the correction and addition of annotations to complete the nucleus labelling process (Protocol 4).

Protocol 4: Custom human-in-the-loop data annotation process using Cellpose 2D followed by generation of binary masks for training and testing the StarDist 2D model

1. Load images into the Cellpose image segmentation software.

2. Use the pre-trained Cyto 3 generalist algorithm model to segment the images, focusing on the nuclei as the regions of interest (ROI).

3. Have a human operator manually correct the annotations made by Cyto 3 to ensure accuracy.

4. Save the annotated ROIs in the '.zip' file format, required by the image analysis software FIJI (ImageJ).

5. Open the images and their respective ROI files in FIJI software.

6. Use a plugin developed by the Laboratory for Optical and Computational Instrumentation (LOCI) at the University of Wisconsin to create binary masks (8-bit grayscale TIFF format) from the annotated ROIs. This process was automated in ImageJ software using a macro code (Code structure outline 4).

7. For images where no nuclei are present to annotate, create a blank mask with the same height and width dimensions as the original image.

8. An ImageJ macro code script was created to processes image files from a source directory, applies Region of Interest (ROI) masks from another directory, and saves the processed mask images to a destination directory. If no ROI mask exists for an image, it creates a blank binary mask.

However, since the StarDist 2D training process returned errors when faced with blank images that had no nuclei as training or test targets, those 38 images were removed from the dataset. The training targets were newly generated image masks where the annotations were highlighted with solid colours to recognise them as separate unique entities, and the background was represented as black (Figure 3.12.).

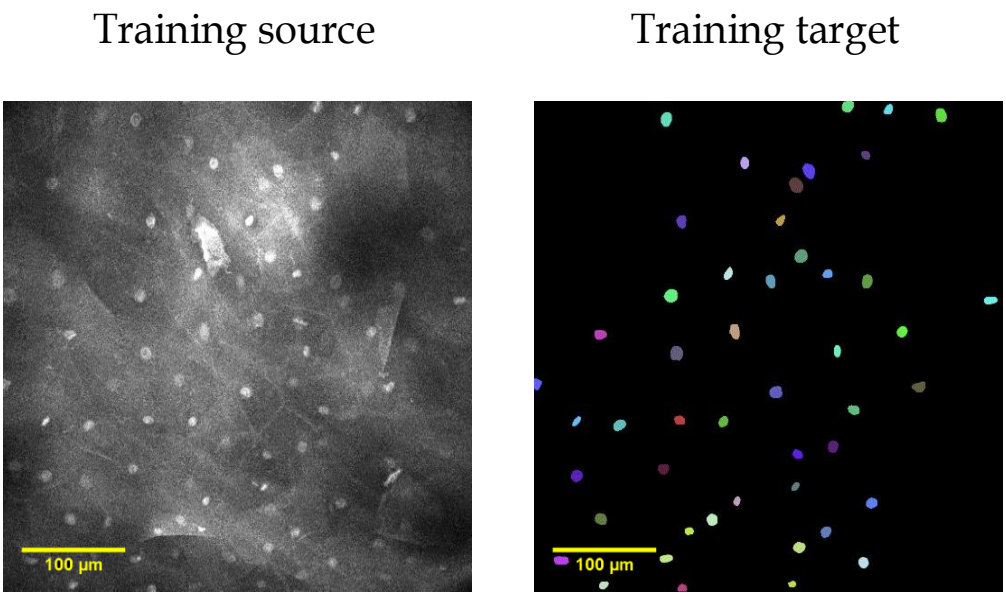


Figure 3.12. An example of a training source image with its binary ROI mask counterpart

Code structure outline 4: Generating binary ROI images from annotations (ImageJ macro code)

1) Prompt user to select the source directory for images.
2) Prompt user to select the destination directory for saving prediction masks.
3) Prompt user to select the directory containing the ROI files.
4) Retrieve a list of all files in the source directory.
5) Loop through each file in the list:
    a) Extract the base name of the current image file.
    b) Construct the full path to the corresponding ROI file.
    c) Construct the full path to the current image file.

d) Print the current file being processed.

6) Check if the current image file exists:
   a) Open the image file.
   b) Store the title of the opened image.
7) Check if the corresponding ROI file exists:
   a) Open the ROI file.
   b) Show all ROIs.
   c) Get the count of ROIs.
8) If there are ROIs:
   a) Apply ROI mapping.
   b) Get the title of the generated mask image.
   c) Save the mask image as a TIFF file in the destination directory.
   d) Close the mask image.
9) If there are no ROIs:
   a) Get the dimensions of the current image.
   b) Create a new blank 8-bit black image with the same dimensions.
   c) Save the blank mask image as a TIFF file in the destination directory.
   d) Close the ROI image.
10) If the ROI file does not exist:
   a) Get the dimensions of the current image.
   b) Create a new blank 8-bit black image with the same dimensions.
   c) Save the blank mask image as a TIFF file in the destination directory.
11) Reset the ROI Manager.
12) Close the original image.
13) If the image file does not exist, print a file not found message.

3.5.2.2. Training and testing the models

The training and evaluation of StarDist-2D was conducted using the ZeroCostDL4Mic platform that is designed to make deep learning more accessible to the microscopy community.

The code was written in user-friendly notebooks that run on Google Colaboratory. Google Colaboratory is a cloud-based Jupyter notebook environment that allows users to write and execute Python code with free access to GPUs making it ideal for machine learning, data science, and research. The StarDist 2D module in ZeroCostDL4Mic is specifically designed to be run online using CPU and GPU processing available on Google Colaboratory. In the current study the Google Colaboratory Pro version was used that provides 83.5 Gb of system RAM and 40 Gb of GPU RAM.

The StarDist notebook sets up the required environment, including installing necessary Python libraries such as TensorFlow, StarDist, with other dependencies such as csbdeep (v 0.8.0), and cuda (v 12.2.140 Build cuda_12.2.r12.2/compiler.33191640_0). A Google Drive containing the training and evaluation images was linked to the Colaboratory notebook. Training included setting hyperparameters, defining the model architecture, and running the training process (Table 3.7.). The initial training weights of the models were imported from the pre-trained 2D_versatile_fluo_from_Stardist_Fiji model which comes with StarDist 2D.

All the default StarDist 2D hyperparameters were used, excluding number of epochs (Table 3.7.). Three variations of 5, 10, and 15 epochs were used to train 3 separate models. Probability threshold is a confidence score threshold where segmentation confidence by the model needs to be above this threshold to be considered as a valid prediction. This helps to filter out low-confidence detections that are likely to be false positives. Among the predicted objects that satisfy the probability threshold, if there is an Intersection over Union (IoU) overlap between two boxes that exceeds a non-maximum suppression (NMS) threshold, the predictions with the lower confidence score will be suppressed (discarded). This tries to ensure that each detected object is represented by only one bounding box. Values for probability threshold and NMS threshold were determined automatically by the models during the optimization step.

Table 3.7. Hyperparameters selected for training the StarDist 2D model

Hyperparameter	Explanation	Default value
Patch size	The size of image patches extracted for training, affecting memory usage and model precision	1024 x 1024
Batch size	The number of image patches processed simultaneously in one training step, impacting speed and stability	2
Percentage validation	The proportion of the dataset reserved for validation to monitor model performance and prevent overfitting	10
N rays	The number of radial rays used to represent star-convex shapes, influencing segmentation accuracy and efficiency	32
Grid parameter	The downsampling factor for the prediction grid, balancing model speed and resolution	2
Initial learning rate	The starting step size for updating model weights, affecting training speed and convergence	0.0003
Probability threshold	The minimum confidence score for accepting a detected object as a valid prediction	0.45
NMS threshold	The non-maximum suppression (NMS) threshold to merge overlapping detections, preventing duplicate predictions	0.3

3.5.2.3. Performance assessment

After training, the module included steps to evaluate the performance of the model. This involved metrics such as Intersection over Union (IoU), precision, F1 score, and panoptic quality (PQ). IoU measures the pixel overlap of predicted objects over the ground truth targets defined during annotation. The IoU threshold for this study was at the default value of 0.5. The IoU scores were expressed in terms of the mean true score (MTS) and mean matched

score (MMS) in order to gain a more nuanced understanding of the results. MTS quantifies how well the overall predicted segments overlap the ground truth segments by taking the mean of the IoUs across an entire image. MMS measure the average IoU score of predicted objects that have a corresponding matched ground truth object (IoU threshold >0.5). PQ is a comprehensive metric (similar to F1 score) that evaluates both the detection and segmentation quality of the model (Equation 1).

$$PQ = \frac{\sum_{(p,g)} IoU(p,g)}{TP + \frac{1}{2}FP + \frac{1}{2}FN}$$

Equation 1: Calculation of panoptic quality (PQ)

In Equation 1, p is a predicted object, g is a target annotated object, IoU is Intersection over Union, TP is True Positives, FP is False Positives and FN is False Negatives. These metrics across all 3 models were compared using summary statistics, and box plots. The Shapiro-Wilk test for normality of data distribution was conducted. A paired sample one-way ANOVA was carried out to analyse the differences between the models.

3.5.3. Experiment 3 – Machine learning diagnostic analysis of segmented features

This experiment involves the use of measurements of the size, shape, fluorescence intensity and distance analysis of segmented nuclei as input features for machine learning classification models to predict OED and OSCC.

Objectives:

1. To utilise the dataset of nuclei measurements for ML diagnostic triage analysis
2. To train, fine tune and optimise the ML models trained on these different versions of the measurement data.
3. To evaluate and compare the performance of all ML models across the different data analysis approaches.

3.5.3.1. Measurements

The best performing model from the StarDist training experiment (Experiment #2) was chosen to segment nuclei in a dataset composed of both acriflavine and fluorescein augmented in vivo confocal micrographs across the 4 diagnostic triage categories of 'no dysplasia', 'lichenoid', 'low-risk', and 'high-risk'. Lichenoid encompassed all oral lichen planus and oral lichenoid inflammation cases. Low-risk contained low-grade dysplasia cases, and High-risk represented the high-grade dysplasia and OSCC cases. Measurements of the size, shape, fluorescence intensity and distance analysis of segmented nuclei were used are features for machine learning classification models to predict OPMDs and OSCC (Table 3.2.).

Confocal micrographs captured from 54 participants possessing lesions represented at least one of the 4 diagnostic triage categories were included for this experiment.

The nucleus measurements collected in Fiji (ImageJ) for each nucleus are described in Table 3.8.

Table 3.8. Nuclei measurements as features for machine learning diagnostic analysis

Measurement	Explanation
Area	Region of selection in square pixels
Mean pixel intensity	Average pixel brightness/pixel intensity value (0-255) within the selection.
Pixel intensity standard deviation	Standard deviation of pixel brightness/pixel intensity values.
Circularity	A value of 1.0 indicates a perfect circle. As the value approaches 0.0, it indicates an increasingly elongated shape. *Formula:* $4\pi * \frac{area}{perimeter^2}$
Aspect ratio (AR):	Ratio of length of major axis to length of minor axis (inverse of circularity)
Integrated density	The sum of the values of the pixels in the image or selection

These images were loaded onto a private secure Google Drive and loaded into the StarDist 2D ZeroCostDL4Mic notebook (von Chamier et al., 2021). The protocol for nucleus measurement was as follows:

Protocol 5: Computing nucleus measurements on CLE images using StarDist 2D and Fiji (ImageJ)

1) Running the StarDist 2D Model on the Dataset
2) Accessing ZeroCostDL4Mic Notebook:
    a) The ZeroCostDL4Mic Google Colaboratory notebook was accessed in a browser.
    b) The runtime was set to use a GPU, specifically the NVIDIA Tesla T4 GPU.
3) The dataset of images was uploaded to the Colaboratory environment.
4) The pretrained StarDist 2D model was loaded and run on the dataset to generate predictions.
5) Once the model had processed the images, the predicted Regions of Interest (ROIs) were downloaded as .zip files.
6) Fiji was opened, and measurements were set via Analyse > Set Measurements by checking the following options:
7) A new macro was created in Fiji for loading the images and their predicted ROIs and calculating nucleus measurements. The measurements were then as saved comma separated values files (Code structure outline 5).

Code structure outline 5: Loading images and their corresponding ROIs to make nucleus measurements and store them in comma separated value files (ImageJ macro code)

1) Prompt user to select directory containing original images
2) Prompt user to select directory containing ROI files
3) Prompt user to select output directory for results
4) Get list of image files in the original images directory
5) For each image file in the list:

a) If the file is a .tif image:

(i) Open the image file

(ii) Extract the image title

(iii) Construct the path to the corresponding ROI file

b) If the ROI file exists:

(i) Open ROI Manager

(ii) Load the ROI file into ROI Manager

(iii) Set measurement parameters (area, mean, standard deviation, etc.)

(iv) Measure the image with the loaded ROIs

(v) Save the measurements as a .csv file in the output directory

6) Reset ROI Manager

7) Close the image and ROI Manager

A new Python environment was created which included a Pandas library to collate summary statistics for all nuclei in every dataset image for further statistical analysis (Code structure outline 6).

Code structure outline 6: Collating summary statistics for further analysis (Python code)

1) Define the directory containing the CSV files with measurements
2) Define the output file for summary statistics
3) Initialize an empty Pandas DataFrame to hold all measurements
4) For each CSV file in the directory:
    a) If the file name ends with '_measurements.csv':
        (i) Read the CSV file into a DataFrame
        (ii) Append the data to the DataFrame holding all measurements
        (iii) Calculate summary statistics (mean, standard deviation, etc.) for each measurement type
        (iv) Save the summary statistics to a CSV file

3.5.3.2. Data Preprocessing and Normalisation

All the nuclei measurements for confocal micrographs belonging to both contrast agents were collected in MS Excel worksheets (version 2410, Microsoft, U.S.A.). Prior to ML analysis, all object measurements were normalised using z-score standardisation. The method (scikit-learn tool 'StandardScaler') transforms the data so that it has a mean of 0 and standard deviation of 1 (Equation 2).

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Equation 2: Z-score standardisation formula

Where X_i is the original value of the feature, μ is the mean of the feature, σ is the standard deviation of the feature, and Z_i is the transformed value. This transformation was applied to all measurements of all acriflavine stained nuclei.

This ensured that all six measurement metrics were on the same scale, preventing any single metric (e.g. Integrated density) from disproportionately influencing the classification process. The standardized values were used as input for further analyses.

3.5.3.3. Analysing relationships between measurements

Summary statistics of mean, median, and standard deviation were calculated for all measurements across all diagnostic categories.

Box plots were created for each measurement across all 4 diagnostic categories to visualise the variance of features across the diagnostic categories. Histograms with a line of best fit were plotted for all diagnostic categories individually for each of the measurements to visualise patterns in the data distribution. The relationships between the measurements and the diagnostic classes were examined as well as relationships among the measurements themselves across both contrast agent datasets. These analyses help identify patterns, dependencies, and potential redundancies in the data.

a. Univariate analysis – One-way Analysis of Variance (ANOVA)

A one-way ANOVA was carried out to determine whether the mean measurement values for all the images differ significantly across the diagnostic classes. The null hypothesis for this test was that the mean measurement values are the same across all diagnostic classes. A statistically significant p-value of <0.05 would indicate if that measurement associated with the diagnostic classes may be a candidate for machine learning analysis. This ANOVA test was carried out in python using the 'statsmodels.formula.api' library.

b. Multivariate analysis – Spearman's correlation

A Spearman's correlation was carried out to assess the strength of the relationships between pairs of measurements. Unlike Pearson's correlation, Spearman's correlation does not assume linear relationships or normally distributed data, making it suitable for identifying non-linear dependencies. The results were summarised in a correlation matrix with the strength and direction visualised using a heatmap. Positive correlations indicated that as one measurement increases, the other tends to increase. Negative correlations indicate that as one measurement increases, the other tends to decrease. Strong correlations may suggest redundancy, while weak correlations may indicate independence in terms of using both of those variables as inputs in the feature vector.

3.5.3.4. Machine learning models

The dataset chosen for training and testing the ML model was the same as the data used for the nucleus measurement analysis.

All images containing less than 10 nuclei were removed from the dataset. The images were divided into 80% training and 20% test images by random selection. The features of this dataset involved the 4 diagnostic category classes 'no dysplasia', 'lichenoid', 'low-risk', and 'high-risk' as being the independent variables (Table 3.2.).

In this study 4 ML models were developed for each of the 4 types of ML models: logistic regression, support vector machines, random forest and XGBoost. The ML model development and evaluation followed the exact same steps which are described under Section 3.4.3. Machine learning development. The models were all evaluated and compared based on their test performance using Protocol 3.

The nuclei measurement data underwent the following preprocessing steps:

1. **Encoding:** The categorical variables were encoded using the 'LabelEncoder' class from the sklearn python library. This involved converting the categorical target class labels into numerical form. The purpose of doing so is logistic regression need numerical targets as they are included in mathematical formulae. Additionally, it provides efficiency and better memory management.

2. **Value scaling:** The dependent variable features which constitute the nucleus measurements were standardised using the 'StandardScaler' class from the sklearn python library. This involved ensuring that all features were on the same scale having a mean of 0 and a standard deviation of 1. This improves model performance as it prevents the model from giving undue importance or

punishment to features having larger numerical ranges simply due to their scale (such as integrated density).

3.5.3.5. Approach 1 - Direct classification of measurement means

All 4 types of ML models were directly applied to the mean scores of the nucleus measurements of area, mean pixel intensity, standard deviation of pixel intensity, integrated density, circularity and aspect ratio for each image. The data was treated with Z-score standardisation while being used as model inputs (**Equation 2**).

3.5.3.6. Approach 2 - Clustering and feature selection

This approach involved clustering all nuclei measurements and using the resulting nuclei cluster proportions per image as the information that represents each image for ML model training and testing.

a. Dataset preparation and Feature Representation

Each image in the dataset contained a variable number of objects, with each object described by six distinct measurement metrics. To prepare this data for classification using a classifier algorithm the feature dimensions needed to be reduced to a fixed dimensional vector for each image which would be

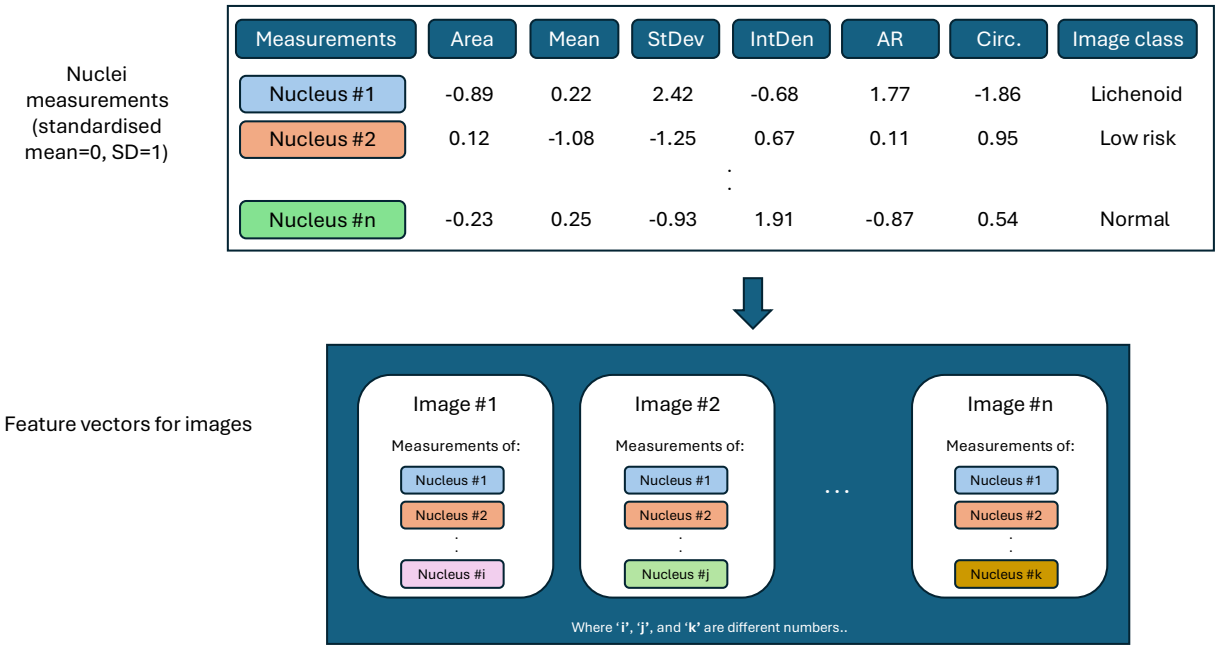


Figure 3.13. Representation of feature standardisation followed by constructing feature vectors with raw measurements as features

independent of number of nuclei measured across each image. The nuclei measurements were represented as six-dimensional feature vectors, with one vector per nucleus. To create a fixed-length feature vector for each image, we employed a k-means clustering approach to group nuclei based on their six-dimensional measurement vectors (Figure 3.13.).

K-means clustering is an unsupervised learning algorithm that partitions datapoints into k distinct clusters based on similarity of features. This algorithm is centroid based, and its goal is to minimise the within-cluster sum of squares (WCSS), which is a measure of how close each point is to its cluster's centroid. The data points are plotted in an n-dimensional space where n is the number of features in the dataset. K-means begins by randomly selecting k initial centroids in this feature space. Each data point is assigned to the nearest centroid and the Euclidian distances between all of these points and their centroids are measured. The algorithm then recalculates and iteratively refines the centroids followed by assigning the datapoints to those clusters.

This method enabled the transformation of variable-length object measurements into a consistent-length representation for subsequent classification tasks (Figure 3.14.).

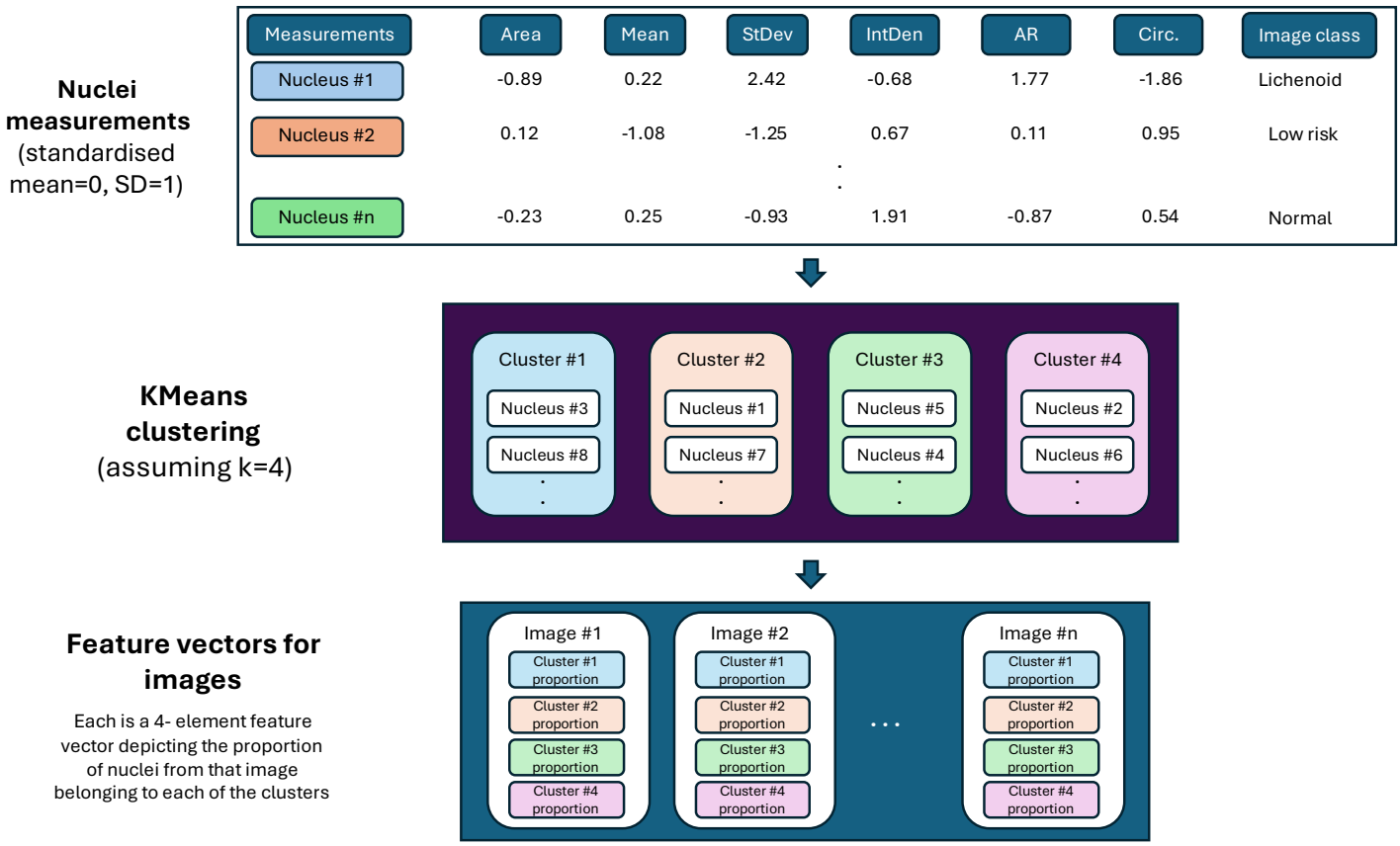


Figure 3.14. Workflow adopted of standardising measurements, K-means clustering all nucleus measurements, and constructing fixed-length feature vectors

b. K-value determination and Feature selection

The input to the k-means algorithm were the six-dimensional feature vectors corresponding to the standardised measurements of individual objects. Selecting an appropriate number of clusters (k) is critical to get optimal clustering results by balancing cluster compactness and model complexity.

The WCSS was computed for a range of k values from 1-20. An elbow plot was graphed using these values. The point at which the rate of decrease in WCSS began to level off, forming an "elbow" in the plot, was selected as the optimal value for k. To account for the impreciseness of selecting a single k value from the graph, a range of five values around the elbow point in the WCSS vs. k plot were chosen as potential candidates for the optimal k for clustering. All of these k value candidates were applied to each of the feature sets to generate several clustering arrangements that were used for their

individual image feature vector construction and eventually diagnostic classification models.

Combining all 6 measurements when clustering the nuclei measurement data may not provide the best results. Feature selection, also known as dimensionality reduction, is the process of identifying and retaining the most relevant features from a dataset. The aim was to improve model performance, reduce computational complexity, and enhance interpretability by focusing on the features that can effectively differentiate the datapoints while eliminating irrelevant ones.

To identify non-linear relationships between the measurements across all nuclei, Spearman correlation matrices were plotted using heatmaps for each contrast agent dataset. Features with strong positive or negative correlations tend to provide similar information, which can introduce redundancy and noise in the data. Such correlated features do not contribute significantly to separating the data into distinct groups. Correlation analysis helps mitigate multicollinearity, where multiple features are correlated, making it difficult to determine the individual impact of each variable on the outputs of classification models such as support vector machines (SVM) or logistic regression.

Initially, all 6 features (measurements) were tested as inputs for the ML models. Subsequently, features were eliminated sequentially in stages to reduce the dimensionality of the data. The choice of features to eliminate at each stage was determined using the correlation matrix. Feature pairs with the highest correlation value were considered for elimination. For each pair, the sum of their correlations with the remaining features was calculated. The feature with the higher sum value, indicating a higher overall correlation with the rest, was deemed likely to provide noisy or redundant information and was eliminated. The remaining features were subsequently combined for the clustering algorithm. This process was repeated three times, reducing the 6 features to 3 features. Each of the feature sets (6, 5, 4, and 3) were used for all k-value clustering and diagnostic classifier model development.

c. Image Feature Vector Construction

For each feature set, k-means clustering was applied to all nucleus measurements across the entire dataset, ignoring their image membership. Each nucleus was assigned to one of the k clusters based on its similarity to other objects in the dataset.

For each image, a fixed-length feature vector was constructed by computing the proportion of nuclei assigned to each of the k clusters. This approach captured the relative distribution of objects across the clusters, rather than simply counting the absolute number of objects (Equation 3).

$$v = [\frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_k}{N}]$$

Equation 3: Vector representation of the k-dimensional feature vector

where n_k represents the number of objects in the image assigned to the k^{th} cluster, and N is the total number of objects in that image. This proportional representation ensured that the feature vectors reflected the relative composition of objects within an image, making them less sensitive to the absolute number of objects in different images.

With 4 feature sets (i.e. 6,5,4, & 3 features) and 5 k-value candidates resulted in the creation of 20 image feature vector datasets (**Code structure outline 4**).

Code structure outline 4: Clustering all nucleus measurements for different values of k, z-score standardisation, and image feature vector construction to prepare the data for ML analysis

1) Setup and Initialization
    a) Define the directory containing the input files.
    b) Create a list of k_values for clustering and generate corresponding file names dynamically.
    c) Specify column names for entities, clusters, and class labels.
    d) Define Helper Function
    e) Define compute_proportional_features: Accept a group of data and the number of clusters as input.
2) Calculate the proportion of objects in each cluster using value_counts (normalize=True).
    a) Return the proportions as a vector.
3) Process Data for Each k-value
    a) Iterate over each k_value and corresponding file name.
    b) Load the data from the Excel file.

c) Rename the cluster column to a generic name for consistency.

d) Group the data by entity and class columns.

4) Apply compute_proportional_features to calculate proportions for each cluster.

a) Convert the results of the proportions into a Pandas DataFrame.

5) Reset the index to have entity and class as columns.

6) Rename the proportion columns dynamically based on the number of clusters.

7) Z-Score Standardization

a) Import StandardScaler from sklearn.preprocessing.

b) Standardize all cluster proportion columns using StandardScaler.

8) Save Processed Data

9) Save the standardized cluster proportions to a new Excel file for each k-value.

10) Output Completion Message

11) Print a message indicating that all files have been processed and saved.

d. Classification of Images

Once the proportional feature vectors were generated, they were used as inputs for supervised classification models to predict the diagnostic class of each image among ‘no dysplasia’, ‘lichenoid’, ‘low-risk’ and ‘high-risk’ categories (Table 3.2.).

The distribution of images between training and test groups was undertaken using an 80%-20% split. All classification models underwent grid search hyperparameter optimisation and 5-fold cross validation. The test performance of all models was compared using **Protocol 3**.

3.5.3.7. Approach 3 - Euclidean distance-based features

This method involved identifying the centroids of all segmented nuclei and measuring all unique Euclidian distances between these nuclei centroids as features for machine learning classification models.

a. Distance measurements

Following the StarDist segmentation process on the entire acriflavine and fluorescein datasets, the centroids of all the segmented nuclei were identified using ImageJ (Fiji).

1. The X and Y co-ordinates of the centroid of each nucleus across all the images were identified by ImageJ and stored in a data table (file type = .csv).
2. The '*pdist*' function from the SciPy spatial distance library was utilised in python to produce a condensed distance matrix containing all pairwise Euclidean distances between all the centroid in the input images (Virtanen et al., 2020).
3. The mean and standard deviation of these distance measurements for all input images were calculated individually and stored together in one data table (file type = .xlsx).

b. Classification

The 4 ML models were trained and tested on this data using the same training-test split as the other approaches. In this approach images with less than 3 nuclei were rejected as there was no way to measure the mean and standard deviation of less than 3 distance measurements.

The mean and standard deviation of Euclidean distance measurements were used as feature inputs for the ML models with the output being classification. The performance of all models was compared using **Protocol 3**.

3.6. Study 4 - Convolutional neural networks with in vivo confocal microscopy for real-time diagnosis of oral cancer and oral potentially malignant disorders

This study describes a protocol for convolutional neural network (CNN) development, training and testing while incorporating hyperparameter optimization, and k-fold cross validation for improved prediction performance. All machine learning studies in this dissertation align with the STARD checklist for reporting diagnostic accuracy and the WHO-ITU checklist for artificial intelligence research in dentistry (Bossuyt et al., 2015; Schwendicke et al., 2021).

Objectives:

1. To develop and train CNN models using the MATLAB's deep network designer for efficient model development, GPU acceleration, and integration with software applications.

1. To develop and train CNN models with hyperparameter optimisation and cross validation using the PyTorch deep learning framework for efficient model development, GPU acceleration, and integration with software applications.

2. To assess the performance of the CNN models developed across both development environments in accurately classifying in vivo confocal microscopy images of oral lesions into lichenoid lesions, OED, and OSCC.

3.6.1. Development of diagnostic triage CNNs in MATLAB

This experiment involved developing and testing deep learning convolutional neural network (CNN) models in the MATLAB mathematics software package (MathWorks, USA) with fluorescence in vivo confocal microscopy images of the oral mucosa for the rapid and accurate detection of OED and OSCC.

a. CNN model development

The diagnostic triage CNN model was constructed using transfer learning in the Deep Network Designer application in MATLAB (MathWorks, USA)

using Inception_v3 as the base template. The CNN development process in this study was identical to the experiment described in Section 3.3.1. Quality filtering CNN development in MATLAB.

Before the images could be loaded into MATLAB to be used by the network, they underwent pre-processing steps using the open-source image analysis software Fiji (ImageJ) using Protocol 2.

For transfer learning two changes were made to the classification head of the model architecture (Figure 3.15.). The last fully connected layer was replaced with a new fully connected layer that has the output classes required for the current study. The final classification output layer was replaced with a new one based on Figure 3.7. that used the cross-entropy loss function to assist in classifying images into the defined diagnostic classes (Table 3.2.).

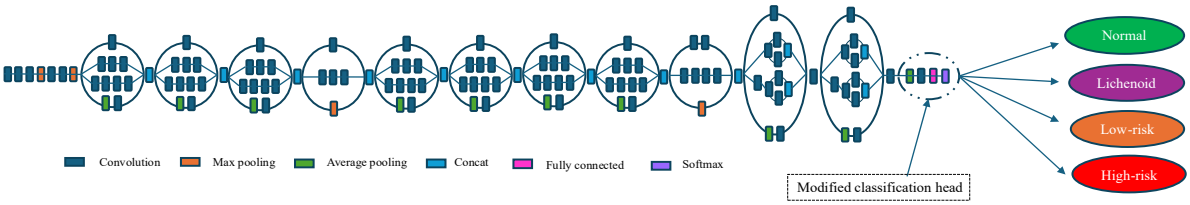


Figure 3.15. Modified Inception_V3 CNN architecture with the diagnostic categories as outputs

For model development 20% of the training images were split for use as validation data during training. Despite the randomness of the training-test split of the dataset it was ensured that images from the same patients were not across both the training and test datasets in an effort to limit overfitting. For data augmentation there was random reflection of the X & Y axes and random rotation of the images from 0-90 degrees along with random rescaling between 1x & 2x times size. The solving algorithm selected for classification was stochastic gradient descent (momentum = 0.9) paired with the cross-entropy loss function. The initial learning rate was set to 0.001, with a learning rate drop factor of 0.1. The training was carried out for 30 epochs in mini batch sizes of 20 images each.

Following training, the CNN was run on the test dataset using uniquely constructed MATLAB programming code. The metrics measured to assess the models were accuracy, sensitivity (recall), specificity, precision and F1 score (Table 3.3.). Additionally, receiver operator characteristic (ROC) curves were plotted and the area under the ROC curve (AUC) was calculated for each diagnostic class vs all for the best performing diagnostic CNN.

In order to create a training and test datasets for the diagnostic triage CNNs, the quality filtering CNN developed in Section 3.3. Study 1 – Micrograph Quality Filtering was applied to all 9168 available images. The resulting diagnostic quality images usable for analysis were used for training and testing the diagnostic triage CNNs models.

The diagnostic triage CNN network was designed to classify the oral mucosal confocal micrographs into the diagnostic categories ‘no dysplasia’, ‘lichenoid’, ‘low-risk’, and ‘high-risk’ as described in Table 3.2.

3.6.2. Development of diagnostic triage CNNs in PyTorch

This experiment involves the development, training, and evaluation of CNNs in the PyTorch deep learning framework utilising hyperparameter optimisation and cross validation in the Python programming language for the diagnostic classification of lichenoid lesions, OED and OSCC in fluorescence human in vivo confocal endomicroscopy images of the oral mucosa.

a. Development framework

The neural networks in this study were all developed within the PyTorch (ver. 2.1.0) framework using Python 3 programming language (ver. 3.11.5) (Paszke et al., 2019; Guido van Rossum & FL Drake, 2021). Python libraries of Numpy (ver. 1.24.3) and Pandas (ver. 2.1.1.) were used for performing numerical calculations and the Sci-kit learn library (ver. 1.3.0.) was used for machine learning applications (McKinney, 2011; Pedregosa et al., 2011a; Van Der Walt et al., 2011). The CNN architecture in this study was a modified Inception_V3 architecture where the final fully connected layer was replaced by a new fully connected layer for the classes required in this study using transfer learning (Figure 14).

The images were pre-processed using Protocol 2 and transferred to PyTorch for specific steps for preparing the data. The PyTorch CNN development in this study was identical to that described in Section 3.3.2. Quality filtering CNN development in PyTorch.

b. Performance assessment

The performance metrics used to assess all CNN models in this study were accuracy (%), sensitivity, specificity, precision and F1 score as described in Table 3.3. The accuracy results across all folds of the 5-fold cross-validation

were averaged to estimate the most optimum hyperparameter combination (**Code structure outline 1**). All trained models were ranked based on an aggregation of ranks for 5 metric scores calculated. The overall rank was calculated by ranking the aggregate rank scores for each hyperparameter combination using **Protocol 3**. All statistical analyses were performed using the Python sci-kit learn library (Pedregosa et al., 2011b).

From the images filtered and deemed to be of diagnostic quality, 80% were randomly categorised as training images and the remaining 20% were kept aside for testing for both contrast agents. Within the training set, 5-fold cross validation was carried out to try 5 different combinations of learning-validation sets that were also divided by a ratio of 80:20.

3.7. Study 5 - Deep learning diagnostic classification for OED and OSCC in a pre-clinical murine model of oral carcinogenesis

This study involved developing and evaluating the performance of deep learning convolutional neural network (CNN) models for the rapid and accurate classification of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) using fluorescence in vivo confocal microscopy imaging in an oral cancer mouse model.

All machine learning studies in this dissertation align with the STARD checklist for reporting diagnostic accuracy and the WHO-ITU checklist for artificial intelligence research in dentistry (Bossuyt et al., 2015; Schwendicke et al., 2021). The protocol for the imaging of mice in this experiment is described in a manuscript by our research group Celentano et al. (2025) (Celentano, Rickard, Low, Silke, Mohammed, Moslemi, Ramani, Franca, Reiner, & McCullough, 2025).

Objectives:

1. Development, training and testing convolutional neural network models using the MATLAB and PyTorch deep learning frameworks on fluorescence in vivo confocal micrographs captured in an oral cancer mouse model for the detection of oral epithelial dysplasia and oral squamous cell carcinoma.

2. To assess the performance of the trained CNN models in accurately classifying in vivo confocal microscopy images of oral lesions from the oral cancer mouse model into the appropriate OED grades (low-grade or high-grade) and OSCC.

3. To compare the performance of the CNN training approaches and identify the most effective strategy for early detection and classification of OED, with the goal of facilitating timely therapeutic intervention and monitoring in pre-clinical models.

3.7.1. Mice

Mice were housed at The Walter and Eliza Hall Institute animal facility under specific pathogen-free, temperature- and humidity-controlled conditions and subjected to a 12 h light/dark cycle with ad libitum feeding. Mice without functional c-REL alleles (*c-Rel*$^{-/-}$) have been described previously (Gerondakis et al., 2006; Köntgen et al., 1995) were generated on a mixed C57BL/6x129SV background and back-crossed onto a C57BL/6 background for >10 generations. Mice lacking TNF (*Tnf*$^{-/-}$) were also backcrossed onto a C57BL/6 background for >10 generations. Control wild-type (C57BL/6) mice were housed in adjacent boxes within the same room.

3.7.2. 4-NQO induction

Oral cavity lesions were induced as previously described by Ni et al. (2021) (Ni et al., 2021). Briefly, 14-week-old mice were treated with regular drinking water ± 100µg/ml 4-NQO (Sigma) for 14 weeks.

From 14 weeks on, all mice were given regular drinking water ad libitum until necropsy 3-12 weeks post 4-NQO treatment cessation (Figure 3.16.). This was followed by biopsy-histopathology for establishing the ground truth dysplasia grade. Confocal imaging was carried out at weeks 14,16,18,22,24, and 26 which marked when the respective mice were euthanised (Figure 3.16.). This facilitated the capture of different stages of the dysplastic process.

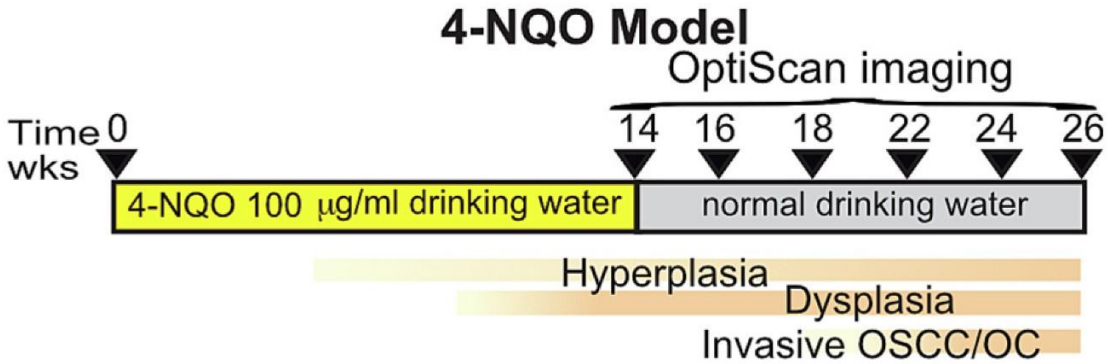


Figure 3.16. Timeline of induction of 4-NQO model in the drinking water of mice (borrowed from Celentano et al., 2025)

3.7.3. Confocal imaging

Confocal images were captured using the hand-held point scanning confocal microscope ViewnVivo FIVE2 (Optiscan Imaging, Australia), equipped with a

3.5 mm diameter x 66 mm long hand-held probe (Figure 3.17.). The imaging was carried out by a trained oral medicine specialist from the Melbourne Dental School, University of Melbourne (Victoria, Australia).

This system features a single 488 channel for illumination with an optical slice thickness/resolution of 5.1 µm (axial) and 0.55 µm (lateral). The probe was directly applied to regions of interest within the oral cavity which included lesions developed at the epithelial surface of the tongue (dorsal/ventral) and the buccal mucosa for image acquisition. The ViewnVivo FIVE2 confocal microscope (Optiscan Imaging, Australia) has identical specifications to the InVivage confocal microscope (Optiscan Imaging, Australia) used for the human imaging analysed in all previously noted studies in this dissertation and as described above in Section 3.2.2.

Figure 3.17. ViewnVivo FIVE 2 confocal microscope (Optiscan Imaging, Australia)

The imaging protocol is described under Protocol 6 and by Celentano et al., 2025 (Celentano, Rickard, Low, Silke, Mohammed, Moslemi, Ramani, Franca, Reiner, & McCullough, 2025).

Protocol 6: In vivo fluorescence confocal microscopy imaging in mice

1. **Anaesthesia:** Mice were sedated with medetomidine-midazolam-fentanyl (MMF) (Fleischmann, Jirkof, Henke, Arras, & Cesarovic, 2016)
2. **Assess clinically:** Assess the oral cavity clinically and record photographs of regions of interest using the flash fitted digital camera

3. **Cleanse oral cavity:** Use pre-soaked sterile cotton swabs (saline) to gently remove food debris/biofilm from the entire oral cavity prior to imaging
4. **Apply contrast agent:** Gently paint oral cavity surfaces with micro-application brushes soaked with 0.1% acriflavine contrast agent and incubate for 1 minute followed by removing unbound excess contrast agent with fresh cotton swab.
5. **Confocal microscopy imaging:** ViewnVivo FIVE2 (Optiscan Imaging) confocal microscope imaging was carried out by placing probe in contact with the oral mucosa to capture images along a z-stack up to depths of 400um of the oral epithelium.

The haematoxylin & eosin stained sections of tissue biopsied from the mice were examined by an independent pathologist who used the binary classification criteria of OED grading (Kujan et al., 2006). This histopathology diagnosis was used as reference to assess performance of CNNs developed on the confocal microscopy images.

3.7.4. Data annotation and pre-processing

Confocal micrographs were labelled for training and evaluating CNNs based on the binary dysplasia pathology scoring of H&E samples from each oral epithelium lesion categorised into ‘no dysplasia, ‘low-grade dysplasia and ‘high-grade dysplasia and oral squamous cell carcinoma’. All micrographs were modified to RGB colour images, down sampled to 299 x 299 pixels using Bicubic interpolation to fit the input criteria for Inception_v3. The 2028 images were split using the ratio 80:20 and allocated randomly into the training (n= 1622) and test datasets (n= 405).

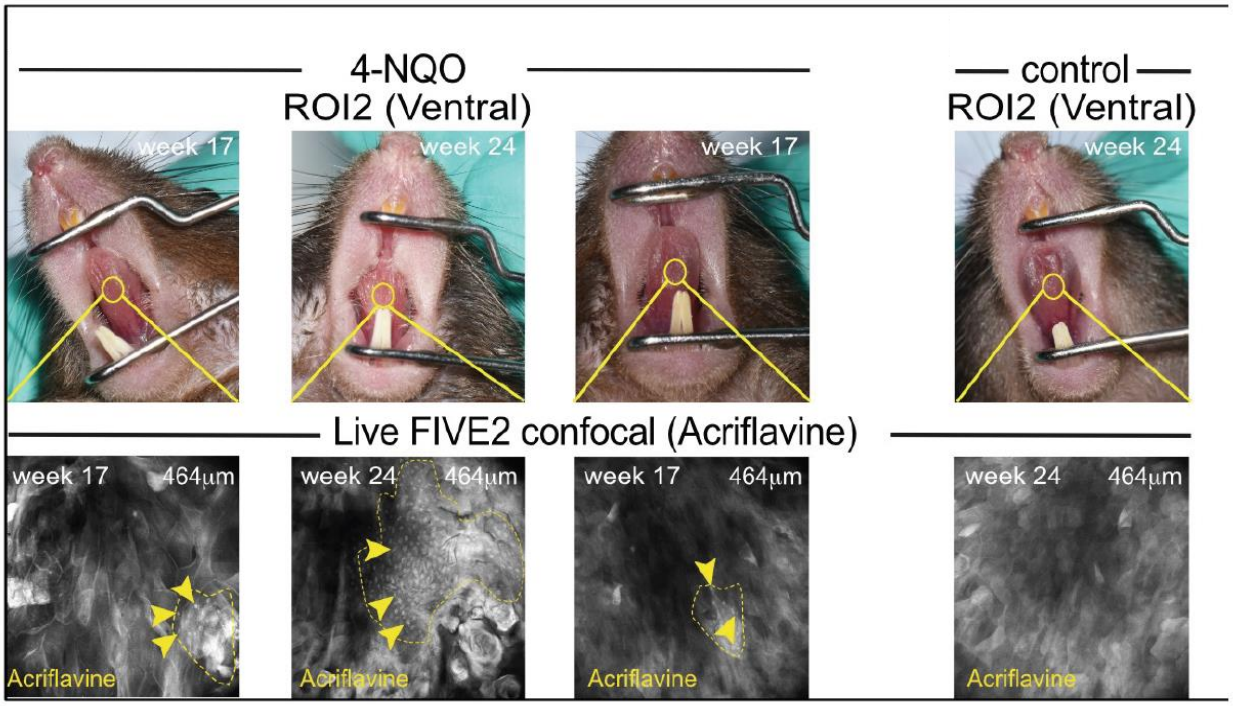


Figure 3.18. Examples of imaging of mice tongue stained with acriflavine using the ViewnVivo (Five2, Optiscan imaging) confocal microscope (Celentano et al., 2025)

3.7.5. CNN development

CNN architecture was a modified Inception_V3 architecture where the final fully connected layer was replaced by a new fully connected layer for the 3 classes in this study using transfer learning.

Fluorescence confocal micrographs were used as inputs for the model with the outputs being diagnostic categories. The image data was divided into three classes based on the binary classification of OED (Kujan et al., 2006):

1) **No dysplasia tissue** (including hyperplasia)

2) **Low-grade dysplasia** (including mild dysplasia)

3) **High-grade dysplasia** (including moderate dysplasia, severe dysplasia, and invasive carcinoma).

This approach aimed to establish a baseline CNN model that could differentiate between no dysplasia, low-grade, and high-grade OED.

Neural networks in this study were developed within MATLAB (MathWorks, U.S.A.) and the PyTorch (ver. 2.1.0) framework using Python 3 programming language (ver. 3.11.5) (Paszke et al., 2019; G Van Rossum & FL Drake, 2021).

Python libraries of Numpy (ver. 1.24.3) and Pandas (ver. 2.1.1.) were used for performing numerical calculations and the Sci-kit learn library (ver. 1.3.0.) was used for machine learning applications (McKinney, 2011; Pedregosa et al., 2011b; Van Der Walt et al., 2011).

The MATLAB diagnostic CNN was developed within the Deep Network Designed application using the steps described in the study in Section 3.3.1., with Protocol 2 for pre-processing, Figure 3.7. for transfer learning, and Figure 3.8. for CNN development. The PyTorch diagnostic CNNs developed in using the steps described in Section 3.3.2. with Protocol 2 for pre-processing, Figure 3.8. for CNN development, Protocol 3 for performance comparison across all individual models across all hyperparameters and cross validation folds, and Code structure outline 1.

4. QUALITY FILTERING MICROGRAPHS

4.1. Introduction

In microscopy, the accuracy of image analysis is crucial for obtaining reliable scientific conclusions, particularly in diagnosis. However, unwanted distortions or anomalies in the form of artefacts can compromise data integrity and lead to misleading results.

Common artefacts include excessive brightness (blooming), pixelation, and noise, which can stem from the physical and technical limitations of microscopy. For example, blooming occurs when a sensor's charge capacity is exceeded, causing excess charge to spill into adjacent pixels, blurring image details. Pixelation arises when magnification exceeds the sensor's resolution, resulting in blocky, indistinct visuals. Similarly, temporal noise are caused by fluctuations in the signal during image acquisition, and motion artifacts due to sample motion during sensor exposure, and these can obscure important structural details within a sample (Roels et al., 2016).

In vivo imaging inside the oral cavity introduces a new layer of challenge in stabilisation of the imaging sensor and the participant being imaged during the acquisition process. Additionally, human saliva, that is occasionally populated with debris, forms a film that covers all tissues to be imaged and could interfere with the microscopy lens. While one solution could involve optimising the image acquisition process, it creates the need for large scale upskilling, technical expertise, and standardisation of imaging protocols while ignoring the random variations introduced by the human participant. The challenge of image artifacts becomes even more significant in high-throughput microscopy, where large volumes of image data are generated at rapid rates. Manually reviewing such vast datasets is impractical, making automated quality control essential. Filtering out images with major artefacts before analysis not only preserves data integrity, but also saves valuable time for microscopists and clinicians, allowing them to focus on interpreting meaningful results rather than sorting through unusable images (Bray, Fraser, Hasaka, & Carpenter, 2012).

To address these challenges, machine learning techniques are increasingly being employed to detect and classify artefacts in high-content screening image data. Advanced quality control workflows use AI-driven algorithms to assess image quality at a cellular level, automatically identifying and excluding problematic images (Litjens et al., 2017). These approaches significantly enhance the reliability of downstream analyses, reducing the risk of errors introduced by poor-quality data.

By integrating automated quality control into high-throughput microscopy workflows, researchers can ensure that only high-quality images contribute to their analyses. This not only improves the accuracy and reproducibility of scientific findings but also streamlines data analysis, enabling experts to dedicate more time to meaningful data interpretation rather than labour-intensive manual data curation.

Aim: To develop and evaluate a deep learning model for accurately classifying diagnostic quality in vivo fluorescence confocal microscopy images of the oral mucosa.

4.2. Methods

The primary objective of this study was to create a pipeline that could filter out poor quality confocal micrographs from a large image database, retaining only high-quality images suitable for further analysis. Two experiments were conducted to achieve this goal.

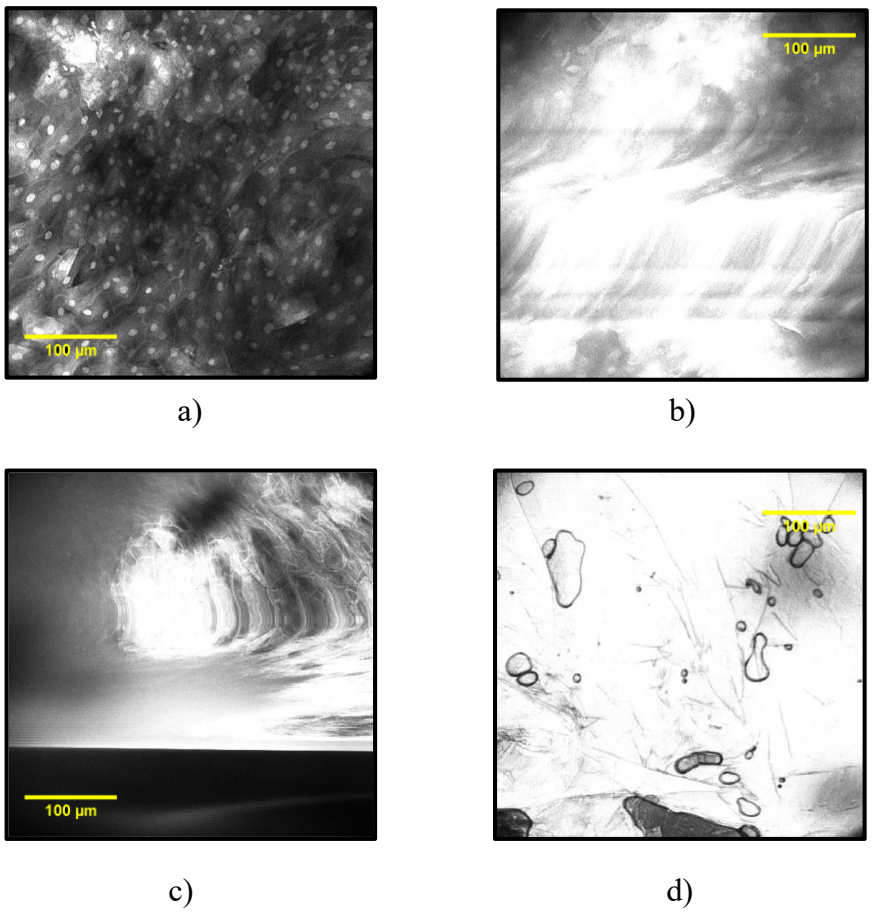
In vivo confocal micrographs analysed in this study were captured from patients with oral mucosal conditions attending the Oral Medicine Department of the Royal Dental Hospital of Melbourne. The in vivo CLE imaging was conducted using the InVivage® (Optiscan Imaging, Victoria Australia). Acriflavine (0.1%) and fluorescein (0.1%) were the topical contrast agents used in this study to enhance the fluorescence imaging contrast of the CLE. Images were captured in the patient's oral cavity from the tongue, buccal mucosa, gingiva & vestibule, floor of mouth, hard palate, and soft palate.

A subset of 1200 images was randomly selected from the entire dataset using the python programming language's 'random.py' library from the entire confocal imaging dataset including 600 images each from each contrast agent to form the training and test datasets for this quality filtering convolutional neural network (CNN) approach. These images were manually annotated by 2 reviewers, one general dental clinician and one oral medicine specialist. Disagreements were addressed by consensus discussion.

The criteria used to determine if an image was of diagnostic quality were (Figure 4.1.):

1. Presence of visible and in focus oral epithelial cell borders or oral epithelial cell nuclei
2. Absence of major artifacts that cover equal to or more than 75% of the field of view of the confocal micrograph
3. Absence of major imaging errors that cover equal to or more than 75% of the field of view of the confocal micrograph
4. Absence of featureless zones that cover equal to or more than 75% of the field of view of the confocal micrograph

CNN models for quality filtering were developed in two different deep learning frameworks of MATLAB (MathWorks, USA) and PyTorch (Paszke et al., 2019) . The MATLAB models were developed using default hyperparameter values for number of epochs and the PyTorch development involved hyperparameter optimisation and cross validation. The quality filtering models developed in this study were named Quality Micrograph Refiner (QMR).



a) b)

c) d)

Figure 4.1. Examples of diagnostic and poor-quality images. a) Diagnostic quality image with clearly visible and in-focus cells and nuclei; b) Poor quality image with excessive imaging brightness and motion blur; c) Poor quality image with featureless zones; d) Poor quality image with excessive brightness and water/saliva droplet artifacts

Experiment 1:

This experiment involved developing a CNN approach for quality filtering in MATLAB software (MathWorks, USA) based on the Inception_v3 architecture. The model, named MATLAB QMR, was developed using the

Deep Network Designer application in MATLAB and trained on 800 manually annotated images and tested on 400 previously unseen images. This entire training and test dataset was comprised of 600 images from each contrast agent dataset.

Experiment 2:

This experiment was conducted in PyTorch and developed a similar CNN model (PyTorch QMR) but employed hyperparameter optimization and cross-validation techniques for improved performance.

Both models were evaluated based on various performance metrics and the best-performing model was then used to filter the entire dataset of fluorescence confocal micrographs, creating a refined dataset of diagnostic quality images for downstream analysis.

The use of default hyperparameters for training CNNs in MATLAB and grid search with 5-fold cross-validation in PyTorch was conducted for exploration of the impact of development environment on model performance. MATLAB is often employed in educational and clinical research contexts where simplicity and rapid prototyping are prioritized, making its well-tuned default settings a practical baseline for typical users. In contrast, PyTorch is a flexible research framework that encourages fine-grained control and methodological rigor, justifying the use of hyperparameter optimization and cross-validation. Practical constraints, such as computational efficiency and ease of experimentation support this approach MATLAB lacks streamlined tools for large-scale grid search and cross-validation, whereas PyTorch integrates seamlessly with such workflows. This design enables a comparative analysis that reveals how much performance improvement is achievable through modern tuning techniques.

CNN performance was assessed using multiple evaluation metrics (Table 2 from Chapter 3):

1. **Accuracy:** The proportion of correctly classified instances out of the total instances.
2. **Sensitivity (Recall):** The proportion of actual positives correctly identified by the model.
3. **Specificity:** The proportion of actual negatives correctly identified by the model.
4. **Precision:** The proportion of predicted positives that are actually positive.

5. **F1-score:** The harmonic mean of precision and recall, balancing both metrics.
6. **Received operator characteristic (ROC) Curve:** A plot of the true positive rate (sensitivity) against the false positive rate across different thresholds.
7. **AUROC (Area Under the ROC Curve):** A single value summarizing the ROC curve, indicating the model's ability to distinguish between classes.

4.3. Quality filtering CNN developed in MATLAB

The resulting quality filtering CNN developed in this study was named 'MATLAB QMR'. A total of 800 images total were randomly selected for testing the QMR model with 400 from each contrast agent dataset. This random selection led to variable number of images from different intra-oral locations (**Table 4.1.**).

Table 4.1: Distribution of images randomly selected for developing the quality filtering CNNs

Intra-oral location	Number of images
Buccal mucosa	300
Floor of mouth	156
Gingiva & Vestibule	234
Hard palate	147
Soft palate	63
Tongue	300

The bespoke modified Inception_V3 model developed in MATLAB's Deep Network Designer application for quality filtering was trained for 960 iterations over 30 epochs in mini batches of 20 images (32 iterations per epoch).

A stochastic gradient descent with momentum algorithm was used to solve for the cross-entropy loss function to update and improve model performance during the training phase. This training process was validated every 10 iterations against the 160 internal validation images. The trained model classified 400 test images into 167 'usable' and 233 'unusable' predictions in 34.90 seconds (0.087 seconds per image) (Table 4.2.).

Table 4.2. Confusion matrix of the performance of the MATLAB QMR on test images (n=400)

		Actual	
		Diagnostic quality	**Poor quality**
Predicted	**Diagnostic quality**	135 (TP)	32 (FP)
	Poor quality	19 (FN)	214 (TN)

The overall accuracy of the model was 87.25% with an error rate of 12.75%. The sensitivity of identifying usable images was 0.88 with the specificity being 0.87, the PPV was 0.81 with the NPV being 0.92. The test result

Acriflavine		
	Actual Usable	Actual Unusable
Predicted Usable	80	9
Predicted Unusable	10	101

a)

Fluorescein		
	Actual Usable	Actual Unusable
Predicted Usable	55	23
Predicted Unusable	9	113

b)

Buccal mucosa		
	Actual Usable	Actual Unusable
Predicted Usable	43	7
Predicted Unusable	7	43

c)

Floor of mouth		
	Actual Usable	Actual Unusable
Predicted Usable	30	2
Predicted Unusable	6	26

d)

Gingiva & Vestibule		
	Actual Usable	Actual Unusable
Predicted Usable	10	1
Predicted Unusable	4	31

e)

Hard palate		
	Actual Usable	Actual Unusable
Predicted Usable	17	9
Predicted Unusable	2	53

f)

Soft palate		
	Actual Usable	Actual Unusable
Predicted Usable	5	0
Predicted Unusable	0	5

g)

Tongue		
	Actual Usable	Actual Unusable
Predicted Usable	30	13
Predicted Unusable	1	56

h)

Figure 4.2. Confusion matrices for the test predictions of the MATLAB QMR subdivided for each contrast agent: a) acriflavine & b) fluorescein and each intra-oral location: c) buccal mucosa, d) floor of the mouth, e) gingiva & vestibule, f) hard palate, g) soft palate, and h) tongue

considerations for the two contrast agents and 6 intraoral location groups (across both contrast agent datasets) are presented Table 4.3.

The MATLAB QMR classified the acriflavine stained images (F1 score = 0.89) better as compared to the fluorescein images (F1 score = 0.77). The model had variable classification performance based on intra-oral site. It achieved a perfect score for all metrics on soft palate images, although it is important to note the soft palate dataset only consisted of 63 images (Table 4.1.).

Besides this, the MATLAB QMR performed best on the floor of the mouth, the buccal mucosa, the tongue (Figure 4.2., Table 4.3.). The model performed worst on the hard palate images (0.76) with gingival & vestibule images showing a slightly better score (F1 score = 0.80). The specificity was higher than the sensitivity for the floor of mouth and gingiva & vestibule images indicating the model was better at identifying diagnostic quality images that did not belong to these groups (Table 4.3.). On the other hand, the sensitivity was higher than the specificity of hard palate and tongue images which indicated the model was better at identifying diagnostic quality images that belonged to these groups compared to the rest (Table 4.3.). The highest model precision was achieved when classifying the soft palate and floor of mouth images (Table 4.3.).

Table 4.3. MATALB QMR test results based on contrast agent and intra-oral location

Test groups		**Accuracy (%)**	**Sensitivity/ Recall**	**Specificity**	**Precision**	**F1 score**
Contrast agents	Acriflavine	90.5	0.89	0.92	0.90	0.89
	Fluorescein	84	0.86	0.83	0.71	0.77
Intra-oral locations	Buccal Mucosa	86	0.86	0.86	0.86	0.86
	Floor of mouth	87.3	0.83	0.93	0.94	0.88
	Gingiva & Vestibule	89.13	0.71	0.97	0.91	0.80
	Hard palate	86.42	0.89	0.85	0.65	0.76
	Soft palate	100	1	1	1	1
	Tongue	86	0.97	0.81	0.70	0.81
Full test dataset		87.25	0.88	0.87	0.81	0.84

4.4. Quality filtering CNN developed in PyTorch

There were 90 PyTorch CNN models developed across all hyperparameter combinations (epochs and learning rate) and all cross-validation folds. They were all ranked based on the performance metrics accuracy, sensitivity, specificity, precision and F1 score with an overall aggregate rank score being calculated to determine the best performing model (Table 4.4.).

Table 4.4. Averaged classification performance of PyTorch QMR across all cross-validation folds for all hyperparameter combinations

Epochs	LR	Accuracy	Sens./ Recall	Spec.	Precision	F1 score	Overall rank
5	0.001	64.0%	0.07	1.00	0.95	0.13	12
5	0.01	86.0%	0.75	0.93	0.87	0.80	5
5	0.1	82.3%	0.75	0.87	0.79	0.76	18
10	0.001	73.1%	0.31	0.99	0.97	0.47	11
10	0.01	85.5%	0.69	0.96	0.91	0.78	2
10	0.1	83.1%	0.74	0.89	0.81	0.77	14
15	0.001	78.5%	0.51	0.96	0.88	0.64	17
15	**0.01**	**88.1%**	**0.78**	**0.94**	**0.90**	**0.83**	**1**
15	0.1	83.7%	0.72	0.91	0.84	0.77	12
20	0.001	83.0%	0.62	0.96	0.91	0.74	9
20	0.01	85.1%	0.70	0.95	0.89	0.78	7
20	0.1	83.5%	0.66	0.94	0.88	0.75	15
25	0.001	84.4%	0.68	0.95	0.89	0.77	8
25	0.01	85.2%	0.70	0.94	0.89	0.79	6
25	0.1	83.7%	0.68	0.93	0.87	0.76	16

30	0.001	85.6%	0.72	0.94	0.89	0.79	3
30	0.01	85.3%	0.71	0.95	0.89	0.79	3
30	0.1	83.8%	0.76	0.88	0.81	0.78	9

LR = Learning rate, Sens. = Sensitivity, Spec. = Specificity

The best hyperparameter combination for CNN performance on the test dataset was 15 epochs with a learning rate of 0.01 (Table 4.4.).

The averaged performance metrics of this parameter combination across all 5 cross validation folds were an accuracy of 88.1% (Rank #1) with a sensitivity/recall of 0.78 (Rank #1), specificity of 0.94 (Rank #10.5), precision of 0.90 (Rank #5), and a F1 score of 0.83 (Rank #1). This model had the best overall aggregate ranking (Rank #1) The best individual model from this parameter combination originated from the Fold #0 with an accuracy of 89.5% with a sensitivity/recall of 0.81, specificity of 0.95, precision of 0.91, and a F1 score of 0.86 (Figure 4.3.).

The PyTorch QMR performed slightly better with the acriflavine dataset (F1 score = 0.86) compared to the fluorescein images (F1 score = 0.85) (Table 5). The results of testing the best PyTorch QMR model were extracted for each intraoral site imaged and the highest F1 scores were for the gingiva & vestibule, and floor of the mouth micrographs and the poorest identification F1 score was for the hard palate images (Figure 4.3., Table 4.5.).

Acriflavine		
	Actual Usable	Actual Unusable
Predicted Usable	74	8
Predicted Unusable	16	102

a)

Fluorescein		
	Actual Usable	Actual Unusable
Predicted Usable	50	4
Predicted Unusable	14	132

b)

Buccal mucosa		
	Actual Usable	Actual Unusable
Predicted Usable	39	3
Predicted Unusable	11	47

c)

Floor of mouth		
	Actual Usable	Actual Unusable
Predicted Usable	30	1
Predicted Unusable	5	27

d)

Gingiva & Vestibule		
	Actual Usable	Actual Unusable
Predicted Usable	13	2
Predicted Unusable	1	30

e)

Hard palate		
	Actual Usable	Actual Unusable
Predicted Usable	12	3
Predicted Unusable	7	59

f)

Soft palate		
	Actual Usable	Actual Unusable
Predicted Usable	4	0
Predicted Unusable	1	5

g)

Tongue		
	Actual Usable	Actual Unusable
Predicted Usable	26	3
Predicted Unusable	5	66

h)

Figure 4.3. Confusion matrices for test predictions by the best ranked PyTorch QMR subdivided for each contrast agent: a) acriflavine & b) fluorescein and each intra-oral location: c) buccal mucosa, d) floor of the mouth, e) gingiva & vestibule, f) hard palate, g) soft palate, and h) tongue

Table 4.5. PyTorch QMR test results for both contrast agents and all intra-oral locations

	Group	Accuracy (%)	Sensitivity / Recall	Specificity	Precision	F1 score
Contrast agent	Acriflavine	88	0.82	0.93	0.90	0.86
	Fluorescein	91	0.78	0.97	0.93	0.85
Intraoral location	Buccal mucosa	86	0.78	0.94	0.93	0.85
	Floor of mouth	90.48	0.86	0.96	0.97	0.91
	Gingiva & Vestibule	93.48	0.93	0.94	0.87	0.90
	Hard palate	87.65	0.63	0.95	0.80	0.71
	Soft palate	90	0.80	1.00	1.00	0.89
	Tongue	92	0.84	0.96	0.90	0.87
Full test dataset		89.5	0.81	0.95	0.91	0.86

4.5. Comparison of MATLAB QMR and PyTorch QMR

The QMR models were developed using different approaches on the same dataset leading to variable results. The PyTorch QMR had an overall better performance with an F1 score of 0.86 (Table 4.6.).

Table 4.6. Overall comparison of the test performance of the PyTorch QMR and MATLAB QMR

	Epochs	Lr	Accuracy	Sens/ Recall	Spec	Precision	F1 score
MATLAB QMR	**30**	**0.001**	**87.25%**	**0.88**	**0.87**	**0.81**	**0.84**
PyTorch QMR	**15**	**0.01**	**89.5%**	**0.81**	**0.95**	**0.91**	**0.86**

Based on intra-oral location the PyTorch QMR had higher overall accuracy for all intra-oral locations except for soft palate (Figure 4). Additionally, the PyTorch QMR model had a test classification rate of 0.03 seconds per image. In contrast, the MATLAB QMR required 0.09 seconds to classify a single image.

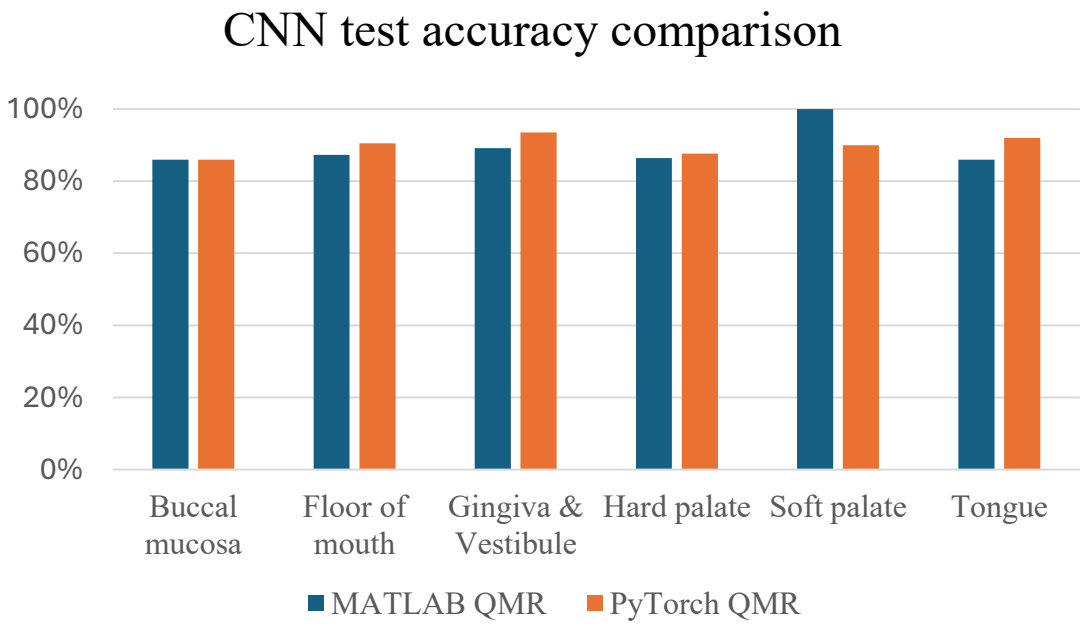


Figure 4.4: Comparison of test accuracy of MATLAB QMR and PyTorch QMR based on intra-oral site

The area under the receiver operator characteristic curve (AUROC) comparison of both the MATLAB and PyTorch QMR showed a marginal difference (Figure 4.5.). The PyTorch QMR had an AUC of 0.94 as compared to 0.92 by the MATLAB QMR based on their performance on the test dataset (Figure 4.5.).

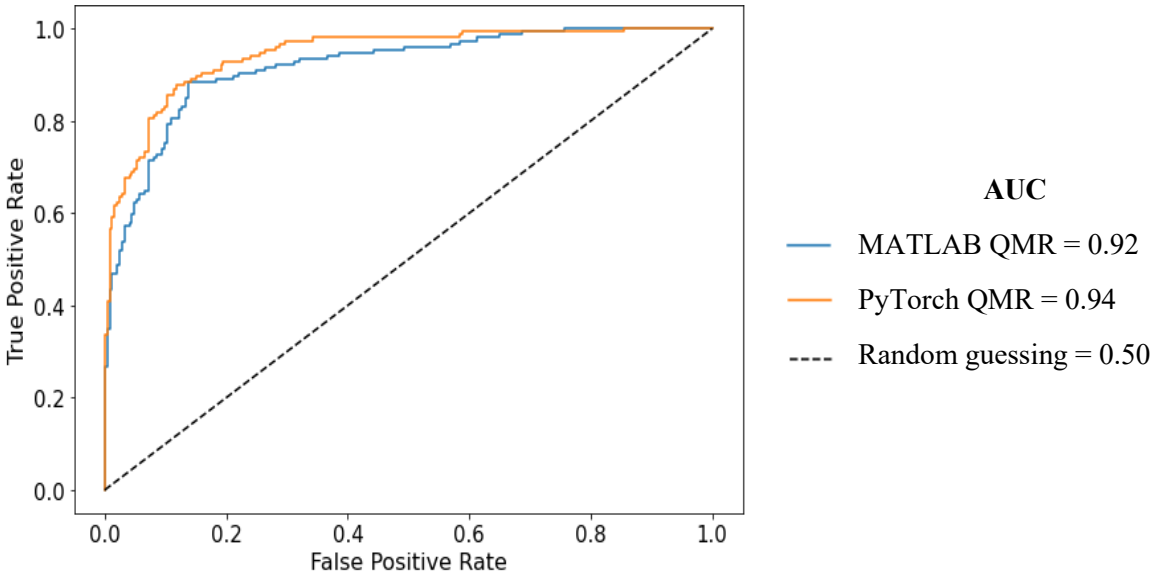


Figure 4.5. AUROC of the MATLAB and PyTorch quality filtering QMR models based on test performance

4.6. Quality filtering the in vivo confocal microscopy dataset

The PyTorch QMR model, being faster and more accurate than the MATLAB QMR model, was chosen to be the definitive quality filtering CNN in this study. It was applied to the entire acriflavine and fluorescein dataset to filter out the diagnostic quality images for all further downstream analysis. Out of a total confocal microscopy databased of 9168 images only 1983 images were identified as being of diagnostic quality (Table 4.7.).

Table 4.7. Diagnostic quality micrographs across all imaging locations after using the QMR on the entire confocal micrograph database

Intra-oral location	Micrographs captured	Diagnostic quality micrographs	Percentage of diagnostic micrographs (%)
Buccal Mucosa	1502	402	26.76
Floor of mouth	891	159	17.85
Gingiva and Vestibule	2757	541	19.62
Hard Palate	230	23	10
Soft palate	150	74	49.33
Tongue	3638	784	21.55
Total	9168	1983	21.63

The proportion of total images which were identified as being of diagnostic quality by the QMR was 21.63% (Table 4.7.). The location with the highest percentage of quality images identified was soft palate (49.33%), with lowest percentage of quality images originated from the hard palate (10%) (Table 4.7.).

The images were randomly divided into 80% training and 20% test image sets for developing the diagnostic triage algorithms (Table 4.8.).

Table 4.8. Image distribution for training and test datasets used to develop the acriflavine and fluorescein diagnostic CNNs

Diagnostic class	Acriflavine		Fluorescein	
	Training	**Test**	**Training**	**Test**
No dysplasia	**449**	**108**	**151**	**35**
Amalgam tattoo	25	6	0	0
Chronic inflammation	16	3	8	2
Denture associated hyperplasia	24	5	4	1
Fibroepithelial polyp	5	1	8	1
Focal papillomatosis	38	9	7	1
Hyperplasia & hyperkeratosis	325	82	119	29
Squamous papilloma	5	1	5	1
Verruciform xanthoma	10	2	0	0
Lichenoid	**275**	**68**	**151**	**37**
Oral lichenoid lesion	175	43	103	25
Oral lichen planus	100	25	48	12
Low-risk	**286**	**69**	**127**	**31**
Atypia	42	10	0	0
Low-grade dysplasia	132	32	91	22
Verrucous hyperplasia	112	27	36	9
High-risk	**71**	**17**	**86**	**22**
High-grade dysplasia	60	15	84	21
OSCC	11	2	2	1
Grand total	**1081**	**262**	**515**	**125**

4.7. Discussion

Intra-oral confocal microscopy has challenges in capturing good quality in vivo confocal microscopy images (Ramani et al., 2023; Yap et al., 2023). Some of the factors influencing the quality of captured confocal micrographs are patient movement while imaging, saliva covering the imaging lens, shaking of the operator's hand while capturing, among others.

Additionally, the contrast agents used for imaging were topically administered on the tissue for a minute before rising with water. This limited exposure could allow for a varied depth of dye penetration into the epithelial surface leading to some images in the z-axis image stack to have image distortions. This leads to the introduction of artifacts and distortions in the confocal micrographs that obscures key landmarks of the oral epithelial architecture and its features.

Thus, the first step in analysing these images was to systematically filter the confocal micrograph images for quality before any diagnostic triage analysis could be conducted. The convolutional neural network (CNN) approach was chosen for this quality filtering task as CNNs are presently state-of-the-art for image recognition tasks. They consist of algorithms with several layers of nodes (called neurons) that can analyse all the pixels in the image while detecting patterns like edges, textures and shapes in different parts of the image (LeCun et al., 2015a).

Instead of developing an entire CNN architecture the present study employed the concept of transfer learning which involves using a pre-trained model (trained on a large dataset like ImageNet) that is adapted to a new but related task by fine-tuning some or all of its layers (Deng, Dong, Socher, Li, Kai, et al., 2009). This allows models to leverage learned features, significantly reducing training time and improving performance, especially when limited data is available (Huh, Agrawal, & Efros, 2016). The Inception_V3 CNN was chosen to be developed using transfer learning in this dissertation due to its exceptionally high performance on the ImageNet dataset with a top-5 accuracy of 93.9%.

CNNs such as Inception_V3 use a stochastic gradient descent (SGD) algorithm for quantifying errors during training and improving performance as they learn the dataset. They do this by looking at the images in the training set multiple times, which is determined by a hyperparameter called 'epochs' and updated their internal parameters such as weights and biases. The direction in which these internal parameters need to be adjusted to improve model performance is determined by a mathematical gradient function that calculates derivatives of all the CNN layers in a hierarchical manner. The magnitude by

which the internal parameter values need to be adjusted is determined using a hyperparameter called ‘learning rate’. These hyperparameters such as epochs and learning rates are determined by the ML practitioner who develops the CNNs, while the internal parameters are learned by the models themselves during training (LeCun et al., 2015a).

The Inception_V3 architecture was modified for quality filtering of the confocal micrographs analysed in this work using two hyperparameter approaches. The first involved development in MATLAB’s Deep Network Designer application where the hyperparameters were chosen empirically. The other approach involved training several modified Inception_V3 models in the PyTorch deep learning framework using the Python programming language while employing hyperparameter optimisation in the form of a grid search algorithm to systematically test 18 different parameter combinations in addition to 5-fold cross validation.

Although the CNN architectures and hyperparameters were kept consistent across MATLAB and PyTorch implementations, differences in performance were observed, which can be attributed to a variety of underlying factors. One major consideration is the variation in backend implementations between the two platforms. PyTorch, for example, uses a highly optimized dynamic computation graph and integrates tightly with low-level libraries that handle operations such as convolution, normalization, and activation functions differently from MATLAB’s Deep Learning Toolbox (Kim, 2017; Paszke et al., 2019). Furthermore, the two environments apply different default settings for key processes such as weight initialization, random seed handling, learning rate scheduling, and regularization techniques, even when the same parameters are manually set.

Data handling and preprocessing pipelines also differ significantly. In PyTorch, image data augmentation, normalization, and batching are typically managed explicitly through ‘torchvision.transforms’ and custom data loaders, offering fine-grained control (Paszke et al., 2019). MATLAB automates much of this process, which can introduce subtle inconsistencies in how input data is prepared and fed into the network (Kim, 2017). Differences in numerical precision (e.g., float32 vs. float64 defaults), hardware utilization (e.g., GPU memory management and parallelization strategies), and multi-threading behaviour can also impact training speed, model convergence, and final inference performance. Lastly, even seemingly minor differences such as library version mismatches, GPU driver updates, or background system processes can contribute to observed discrepancies.

This approach of developing and comparing CNNs developed in MATLAB and PyTorch is justified by its intent to evaluate the trade-off between usability

and performance in two distinct deep learning environments. By training CNNs in MATLAB using default hyperparameters, the experiment mirrors typical usage scenarios and highlights the platform's accessibility. In contrast, the optimized PyTorch models demonstrate the potential performance gains when more effort and resources are invested in model development. The asymmetry in methodology is a deliberate strategy to explore both practical effectiveness and theoretical performance ceilings. It allows for a more comprehensive assessment of each framework's strengths: MATLAB's out-of-the-box convenience versus PyTorch's optimisation capabilities. Moreover, by identifying the extent to which performance can be improved through hyperparameter tuning, the approach provides actionable insights for practitioners deciding between ease-of-use and performance optimization in deep learning workflows.

The developed quality filtering CNNs were named Quality Micrograph Refiner (QMR). Upon ranking all 90 PyTorch QMRs, the best CNN had a higher accuracy (89.5%) and overall F1 score (0.86) compared to the MATLAB QMR (87.25% & 0.84 respectively) across both acriflavine and fluorescein images. Both QMR CNNs had a high area under the receiver operator characteristic curve (AUROC) with the PyTorch QMR have a slightly higher AUC (0.94) compared to the MATLAB QMR (0.92) (Figure 4.4.).

The PyTorch QMR CNN displayed a high classification F1 score on in vivo confocal micrographs taken from the floor of the mouth (0.91), gingiva & vestibule (0.90) and soft palate (0.89). Importantly the sites with the highest incidence of OSCC such as the tongue (0.87) and buccal mucosa (0.85) also displayed a high F1 score. However, the PyTorch QMR performed moderately on hard palate (0.71) images possibly due to the difficulty in accessing imaging sites because of the variation in the shape of the palatal vault. Additionally, the keratinised mucosa of the hard palate could pose challenges with penetration of the incident and reflected confocal microscope laser. A limitation of this model development is the imbalance in the dataset across different intra-oral locations (Table 4.1.). This could indicate the potential limitations of the hardware and CNNs in quality filtering images from highly keratinised tissue.

In addition to being highly accurate, the PyTorch QMR had a rapid classification rate of 0.03 seconds per image which highlights its potential for real-time chairside diagnostic filtering of in vivo captured confocal micrographs. This functionality when used in tandem with diagnostic algorithm can speed up real-time precise diagnostic triage while ensuring the quality standards of imaging are maintained. The quality filtering model once applied to the entire captured confocal microscopy database identified only

21.63% of the images being of diagnostic quality. This highlights the challenges of in vivo imaging with about 1 in 5 captured images being usable for diagnostic triage. Non-keratinised oral mucosa locations such as buccal mucosa (26.76%), floor of the mouth (17.85%), and soft palate (49.33%) showed higher proportion of diagnostic quality images compared to the tongue (21.55%), gingiva (19.62%) and hard palate (10%) (Table 4.7.). This difference in proportion of diagnostic quality images show potential links between tissue keratinisation and diagnostic quality of in vivo confocal microscopy. Further studies to specifically compare diagnostic quality of in vivo confocal micrographs to tissue keratinisation might help refine quality filtering machine learning models such as the one developed in this study.

While this quality filtering approach performed well, future research could focus on enhancing the CNN models through several avenues. Systematic hyperparameter optimization techniques such as Bayesian search algorithms could yield improved convergence and generalization (Abdullah, Hassan, & Mustafa, 2022). Incorporating transfer learning from pre-trained architectures such as ResNet or EfficientNet may offer varied performance outcomes, particularly with limited datasets (He et al., 2016; M. Tan & Le, 2019). Further, aligning low-level implementation details between MATLAB and PyTorch would support reproducibility and clarify framework-specific effects. Architectural enhancements, including neural architecture search or integration of attention mechanisms, may also lead to performance improvements. Additionally, model compression techniques such as quantization or pruning could optimise inference efficiency for deployment. Finally, evaluating model robustness various external datasets would ensure greater real-world applicability and reliability.

While this technology was developed on confocal micrographs the concept of having a quality filtering CNN in tandem with other models for diagnostic triage, could be applied to several different areas of medical imaging. Adopting this methodology could standardise imaging techniques, such as in vivo confocal microscopy.

This quality filtering process can be augmented with a user-feedback system that implements an audio or visual indication so as to indicate that the images being captured are of good or poor quality. This immediate feedback would then assist operators in saving patient time by focusing capture on good quality images. Additionally, this can have implications on training new operators on the recognition of techniques required for quality data capture. This would subsequently be extended to quality control calibration exercises for confocal microscopy operators across a diverse range of clinical and research settings for standardisation of captured image quality.

5. MACHINE LEARNING ANALYSIS OF HUMAN IDENTIFIED QUALITATIVE FEATURES

5.1. Introduction

The early diagnosis of oral squamous cell carcinoma (OSCC) necessitates a combination of conventional expert oral examination and scalpel biopsy. Considering the diagnostic tools available to us presently, the gold standard in diagnosis of malignant lesions and OPMDs is incisional biopsy followed by a histological analysis (M. McCullough, G. Prasad, & C. J. A. d. j. Farah, 2010b). During histological examination, the architecture of the epithelium with distortions in cellular and nuclear structure are analysed. Abnormalities in these observations of the cells in question are referred to as oral epithelial dysplasia (OED) (Pindborg et al., 2012).

Diagnostic triage studies generally use defined diagnostic categories based on the WHO classification for oral epithelial dysplasia (OED). These include the histopathology microscopy signs of OPMDs and OSCC via the identification of abnormal dysplasia variation in nuclear size & shape along with cell size & shape along with hyperchromatism which indicates an increase in nuclear density. The WHO classifies oral epithelial dysplasia (OED) as mild, moderate, and severe dysplasia. Mild dysplasia is cytological or architectural alterations in the lower one-third of the epithelium. Moderate dysplasia relates to epithelial atypia which extends to the spinous epithelial layer at the middle third. Severe dysplasia is represented by epithelial atypia which extends throughout the entire thickness of the epithelium and is sometimes identified as carcinoma-in-situ (Reibel et al., 2017b). Other classification acknowledged by the WHO is the binary OED classification that recognises mild dysplasia as low-grade OED and combines moderate and severe dysplasia into high-grade OED (Kujan et al., 2006). The studies in this dissertation use the binary OED classification.

A review of studies using confocal microscopy for the diagnosis of OED and OSCC evaluated qualitative cellular features and nuclei histopathology landmarks, such as cell size homogeneity, cell crowding, nucleus size homogeneity, nuclear crowding (Ramani et al., 2023). This chapter explores a novel approach to the classification of oral mucosal conditions using in vivo confocal laser endomicroscopy (CLE) coupled with machine learning (ML) algorithms. Using the quality filtered dataset from the quality filtering convolutional neural network study described earlier (Chapter 4), this study represents a step forward in the practical application of ML concepts for diagnostic triage.

The current study utilizes the InVivage® CLE system to capture high-resolution, cellular-level images of various oral mucosal sites. By employing topical contrast agents such as acriflavine and fluorescein.

Central to this study is the evaluation of five critical cellular and nuclear features: cell crowding, cell size homogeneity, nuclei crowding, nuclei size homogeneity, and the presence of fluorescent granules. These features were assessed using both multi-tier and binary classification systems, providing a comprehensive basis for subsequent machine learning analysis. This study involves using different ML models to allow for a robust comparison of different algorithmic strategies in tackling this complex classification problem. The results of this study demonstrate the potential of combining qualitative features of the oral epithelium identified on CLE images with artificial intelligence for the detection of lichenoid lesions, oral epithelial dysplasia, (OED) and oral squamous cell carcinoma (OSCC).

Aim: To develop machine learning models that can classify lichenoid inflammation, oral epithelial dysplasia, and oral squamous cell carcinoma based on qualitative features observed in in vivo captured fluorescence confocal endomicroscopy images of the oral epithelium.

5.2. Methods

In vivo confocal micrographs analysed in this study were captured from 59 patients with oral mucosal conditions attending the Oral Medicine Department of the Royal Dental Hospital of Melbourne. In vivo CLE imaging was collected using the InVivage® (Optiscan Imaging, Victoria Australia). Acriflavine (0.1%) and fluorescein (0.1%) were the topical contrast agents used to enhance the fluorescence imaging contrast of the CLE. Images were captured in the patient's oral cavity from the tongue, buccal mucosa, gingiva & vestibule, floor of mouth, hard palate, and soft palate.

All imaged lesions had a surgical biopsy with standard care histopathology assessment of the site of imaging. All images were categorised based on the matched histopathological assessment into 4 diagnostic classes 'no dysplasia', 'lichenoid', 'low-risk', and 'high-risk' (Table 3.2.). A subset of 600 randomly selected images equally divided between the acriflavine (n=300), and fluorescein (n=300) image sets were taken from the overall dataset that was filtered for quality using the trained quality filtering convolutional neural network (PyTorch CNN) as described in Chapter 4.

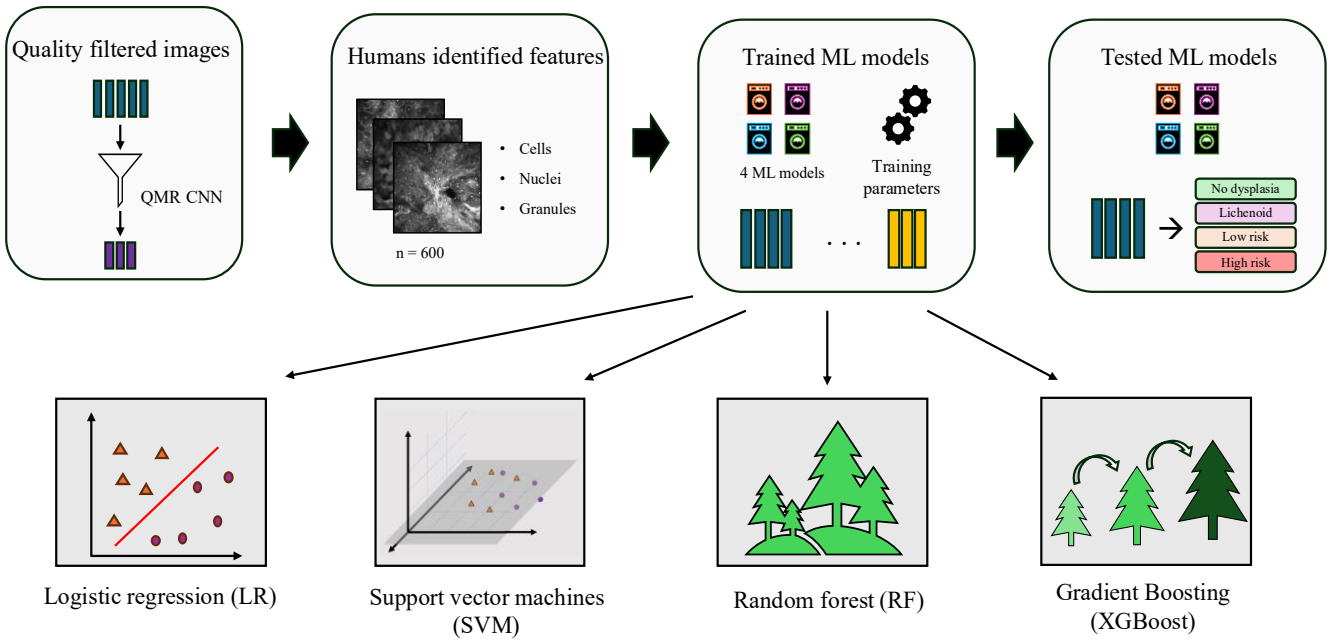


Figure 5.1. Flow-chart for ML diagnostic analysis of qualitative features in in vivo confocal micrographs

The acriflavine and fluorescein images were categorically classified by two oral medicine clinicians to determine the presence and absence of identifiable

cellular and nuclear features (Figure 5.1.). The identified features were (Table 3.4.):

1. Cell crowding
2. Cell size homogeneity
3. Nuclei crowding
4. Nuclei size homogeneity
5. Fluorescent granules

The evaluation of these features was conducted using two distinct systems: a multi-tier system and a binary system.

These 5 selected features represent epithelial cell, nucleus, and cytoplasm characteristics while maintaining relatively low feature dimensionality (Figure 5.1.). Fewer features prevent the ML models from becoming overly complex that could lead to model's overfitting on the training data and showing poor generalisation on evaluation. Feature set A was classified into a multi-tier system as follows:

1. **Cell crowding**:
    a. absent
    b. low
    c. high
2. **Cell size homogeneity**:
    a. regular
    b. borderline irregular
    c. mildly irregular
    d. moderately irregular
    e. highly irregular
3. **Nuclei crowding**:
    a. absent
    b. low
    c. high
4. **Nuclei size homogeneity**:

    a. Regular
    b. borderline irregular
    c. mildly irregular
    d. moderately irregular
    e. highly irregular
5. **Fluorescent granules**:
    a. absent
    b. limited presence
    c. abundant presence

Feature set B was classified into a binary system as follows:

1. **Cell crowding**:
    a. absent
    b. present
2. **Cell size homogeneity**:
    a. regular
    b. irregular
3. **Nuclei crowding**:
    a. absent
    b. present
4. **Nuclei size homogeneity**:
    a. regular
    b. irregular
5. **Fluorescent granules**:
    a. absent
    b. present

Chi-squared analysis was carried out to determine whether there was a statistically significant association between the feature variables and the categorical target variable. Four machine learning (ML) approaches – Logistic regression (LR), support vector machines (SVM), random forest (RF), and XGBoost (XGB), were used to classify qualitative image feature data samples into the four classes (Table 3.2.).

The qualitative categorical features recorded by the 2 investigating oral medicine specialists in two forms as Feature set A and Feature set B were used as categorical input data for the development of these ML models. The features for each image were compiled into a feature vector of 5 elements, with 1 element for each of the 5 features recorded for every single image. These feature vectors were provided to the ML models as input. Each type of ML model processed these feature vectors using their own statistical function and provided a classification output. This output simply contained the diagnostic triage categories described in Table 3.2.

Each ML model was trained using the training subset, optimised using grid search hyperparameter optimisation, and evaluated on the testing subset. The following evaluation metrics were calculated to assess model performance as described in Table 3.3:

- **Accuracy:** The proportion of correct predictions out of total predictions.
- **Sensitivity (Recall):** The ability of the model to correctly identify true positives for each class.
- **Specificity:** The ability of the model to correctly identify true negatives for each class.
- **Precision:** The proportion of true positive predictions out of all positive predictions made by the model.
- **F1 Score:** The harmonic mean of precision and recall, offering a balanced measure of performance.
- **AUROC:** Area under the receiver operator characteristic curve

All analyses were conducted in Python programming language with the scikit-learn, numpy, pandas, and XGBoost libraries.

5.3. Frequency of features across diagnostic categories

A randomly selected subset of 600 images (n=300 per contrast agent) from all usable images filtered out by the quality filtering CNN from Chapter 4 were selected for this study (Table 5.1.). The images in both contrast agent datasets were variably divided between the different qualitative categories for all cellular and nuclear features. Patterns emerged from this distribution which were representative of the diagnostic categories.

Table 5.1..Distribution of the randomly selected set of confocal micrographs for human identified feature ML analysis

Diagnostic category	Acriflavine	Fluorescein
No dysplasia	122	112
Lichenoid	55	79
Low-risk	88	52
High-risk	35	57

5.3.1. Feature set A – Multi-class categorisation

5.3.1.1. Distribution of features in acriflavine feature set A

Cell crowding:

Absence of cell crowding was most frequently seen in No dysplasia, while Low-risk also showed high frequency. All Lichenoid cases showed no cell crowding (Figure 5.2.). Low cell crowding was moderately seen in all classes with no instances in Lichenoid. About 31.25% of the High-risk cases showed high cell crowding. A smaller proportion of low-risk and no dysplasia had high cell crowding (Figure 5.2.).

Cell size homogeneity:

Majority of the no dysplasia, lichenoid and low-risk images showed regular or mildly irregular cell size variation. Several high-risk images also showed regular cell sizes. However, 50% of the high-risk images showed moderate to highly irregular cell sizes (Figure 5.2.).

Nuclei crowding:

Majority of all diagnostic categories showed no nuclei crowding. The highest proportion of nuclei crowding was seen in high-risk with 34.38% of the images showing high levels of nuclei crowding (Figure 5.2.).

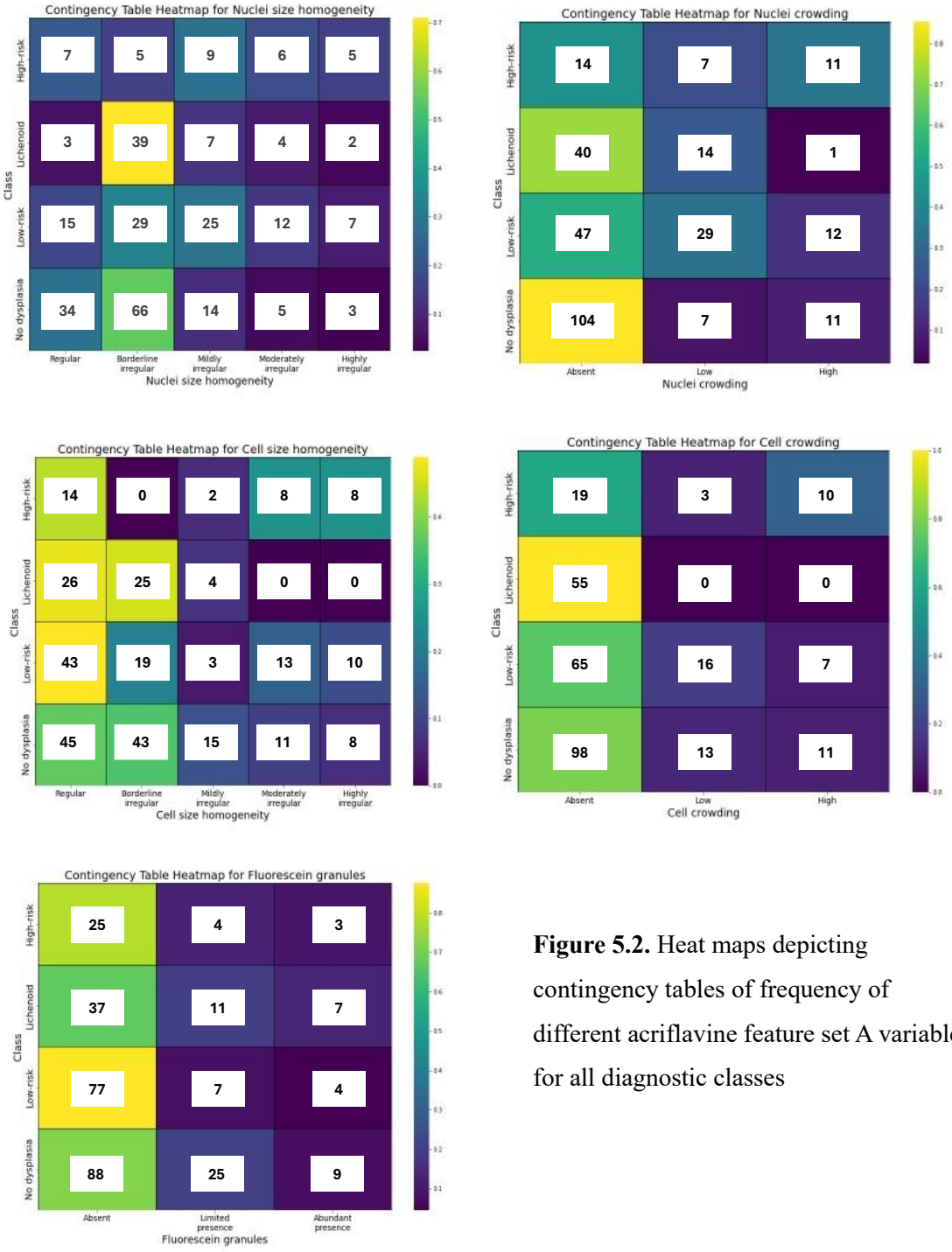


Figure 5.2. Heat maps depicting contingency tables of frequency of different acriflavine feature set A variables for all diagnostic classes

Nuclei size homogeneity:

Mild irregularity in nuclei size was highest frequency seen in no dysplasia, lichenoid and low-risk images. High-risk images were distributed evenly

across all categories with no obvious majority in terms of nuclei size indicating a high variability in nuclei features seen (Figure 5.2.).

Fluorescent granules:

These granules were absent from majority of the images across all diagnostic categories. The next most prevalent category was limited presence with abundant presence being relatively rare across all diagnostic categories (Figure 5.2.).

Overall, these findings suggest that high-risk cases tend to show more irregularities and crowding in cell and nuclei features compared to other diagnostic categories. No dysplasia, lichenoid, and low-risk cases often exhibit similar patterns with less crowding and more regularity in cell and nuclei features.

5.3.1.2. Distribution of features in fluorescein feature set A

Cell crowding:

Absence of cell crowding was predominantly seen in all diagnostic categories, with lichenoid cases showed no cell crowding. The highest proportion of high cell crowding (22.2%) was seen in high-risk cases (Figure 5.3.).

Cell size homogeneity:

Majority of no dysplasia, lichenoid, and low-risk cases showed regular or mildly irregular cell sizes. Even though a large number of high-risk cases showed no cell size irregularity, about 31.48% of high-risk images showed highly irregular cell sizes (Figure 5.3.).

Nuclei crowding:

Majority of the diagnostic categories showed an absence of nuclei crowding. The instances of nuclei crowding across all diagnostic categories were negligible (Figure 5.3.).

Nuclei size homogeneity:

Majority of the images across all diagnostic categories showed regular or mildly irregular nuclei sizes. The highest frequency of highly irregular nuclei sizes was seen in high-risk (75%) (Figure 5.3.).

Fluorescent granules:

Fluorescent granules in fluorescein images were more evenly spread between all diagnostic categories. No dysplasia images showed an even distribution of absence, limited presence and abundant presence of granules. The highest

frequency of no dysplasia, lichenoid, and low-risk images were lacking fluorescence granules.

About 77.78% of high-risk images showed some fluorescent granules (Figure 5.3.). Similar to the acriflavine dataset, these findings in the fluorescein dataset suggest that High-risk cases tend to show more irregularities in cell and nuclei sizes, as well as a higher likelihood of fluorescent granules. No dysplasia, Lichenoid, and Low-risk cases often exhibit similar patterns with less crowding and more regularity in cell and nuclei features. The presence of fluorescent granules appears to be a potentially distinguishing feature for High-risk cases.

Figure 5.3. Heat maps depicting contingency tables of frequency of different fluorescein feature set A variables for all diagnostic classes

5.3.2. Feature set B – Binary categorisation

5.3.2.1. Distribution of features in acriflavine feature set B

Cell crowding:

Majority of the acriflavine images across all diagnostic categories showed no cell crowding. Cell crowding was prominently seen in high-risk images with 40.6% of these lesions showing some amount of cell crowding (Figure 5.4.).

Cell size homogeneity:

Majority of acriflavine images showed regular cells sizes for all diagnostic categories except high-risk lesions. About 56.25% of high-risk cases showed some cell size variability (Figure 5.4.).

Nuclei crowding:

Majority of no dysplasia and lichenoid lesions showed no nuclei crowding. For low-risk lesions the images with and without nuclei crowding seen at about the same frequency. About 56.25% of high-risk cases showed some nuclei crowding (Figure 5.4.).

Nuclei size homogeneity:

Majority of no dysplasia and lichenoid lesions showed homogeneity of nuclei sizes. For low-risk lesions the images with regular and irregular seen at the same frequency. About 62.5% of high-risk cases showed some nuclei size irregularity (Figure 5.4.).

Fluorescent granules:

Majority of all diagnostic categories showed no fluorescent granules. The highest proportion of images with fluorescent granules were seen in no dysplasia cases (48.57%) (Figure 5.4.).

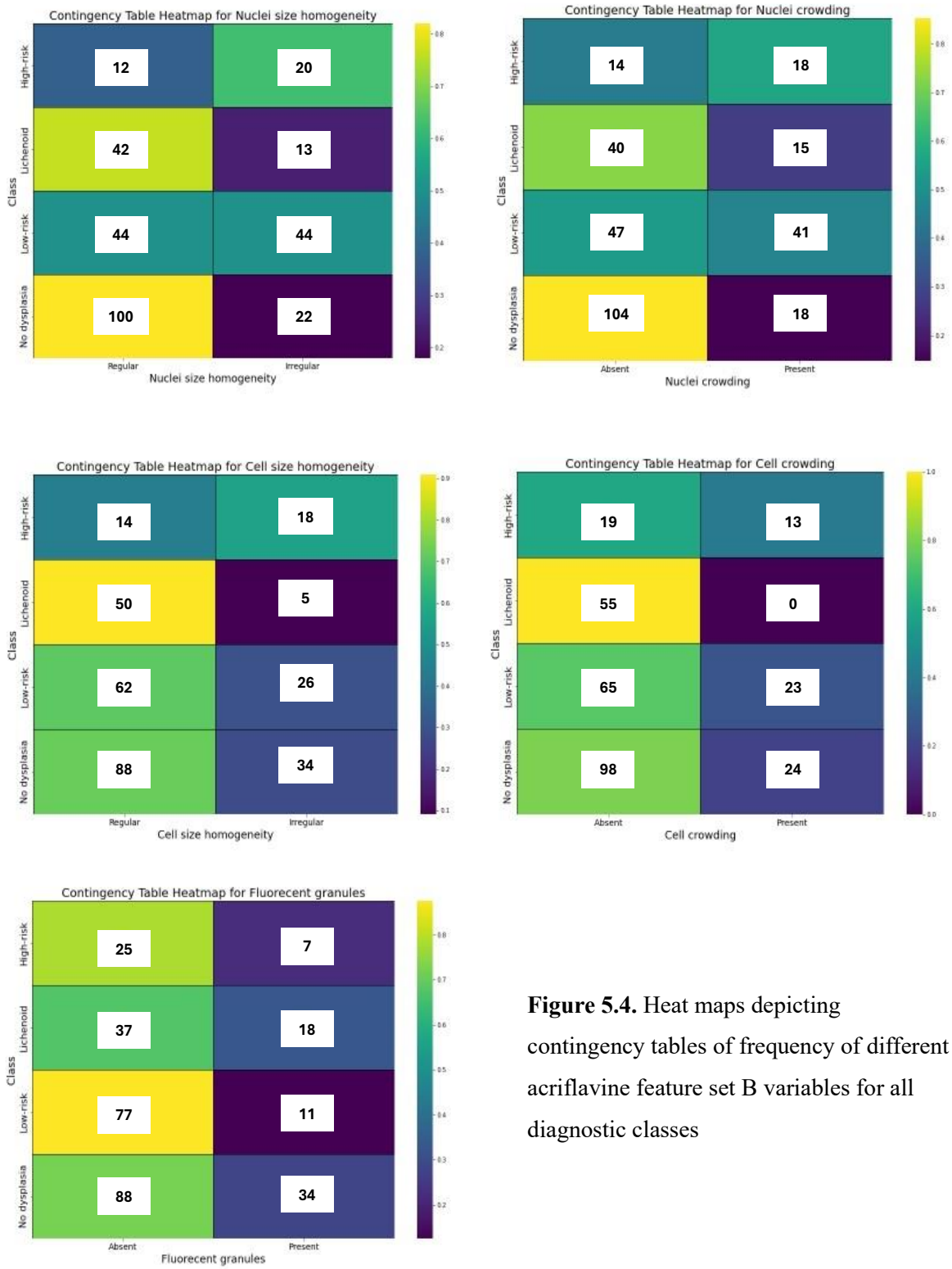


Figure 5.4. Heat maps depicting contingency tables of frequency of different acriflavine feature set B variables for all diagnostic classes

-

Overall, these findings from acriflavine images suggest that High-risk cases tend to show more irregularities in cell and nuclei features, including crowding and size variability. No dysplasia and Lichenoid cases often exhibit similar patterns with less crowding and more regularity in cell and nuclei features. Low-risk cases show some intermediate characteristics.

5.3.2.2. Distribution of features in fluorescein feature set B

Cell crowding:

Majority of images across all diagnostic categories showed no cell crowding. The highest proportion of cell crowding was seen in high-risk (45.09%) images followed by no dysplasia (41.17%) images (Figure 5.5.).

Cell size homogeneity:

Majority of no dysplasia, lichenoid, and low-risk images showed regular cell sizes. The majority of images with any cell size variation was seen in no dysplasia images (46.08%), followed by high-risk images (30.39%) (Figure 5.5.).

Nuclei crowding:

Majority of the fluorescein images showed no nucleus crowing across all diagnostic categories. Only 7% of all fluorescein images showed any nuclei crowding (Figure 5.5.).

Nuclei size homogeneity:

Majority of the fluorescein images showed homogeneity in nuclei sizes across all diagnostic categories. The highest proportion of images with nuclei size variation was seen in no dysplasia cases (49.01%) (Figure 5.5.).

Fluorescent granules:

Majority of no dysplasia, lichenoid and high-risk images showed some fluorescent granules. Low-risk images were close to evenly distributed between having and not having fluorescent granules. The highest proportion of images with fluorescent granules were no dysplasia images (38.29%) followed by lichenoid (24.57%) and high-risk (24%) images (Figure 5.5.).

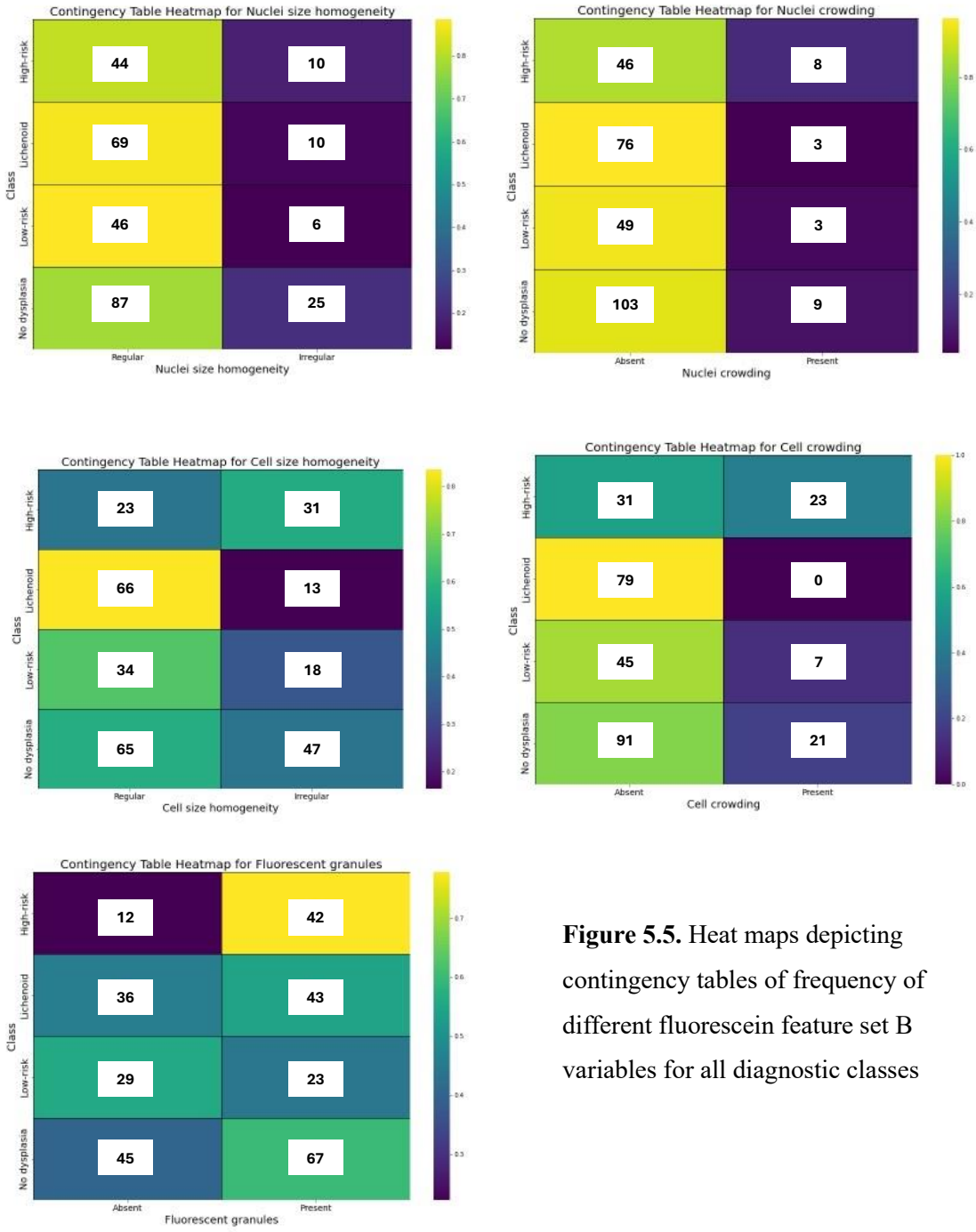


Figure 5.5. Heat maps depicting contingency tables of frequency of different fluorescein feature set B variables for all diagnostic classes

In summary, the findings indicate that high-risk cases generally showed more cell crowding, cell size irregularity, nuclei crowding, and nuclei size irregularity compared to other diagnostic categories. Additionally, the presence of fluorescent granules was more prevalent in no dysplasia and high-risk cases. These findings suggest that these morphological features could be useful in distinguishing between different diagnostic categories.

5.4. Categorical correlation of micrographic features

5.4.1. Feature set A – Multi-class categorisation

The chi-squared test revealed that all features had a statistically significant association with the target variable, as indicated by p-values below the threshold of 0.05 (Table 5.2.).

Table 5.2. Chi-squared test results for feature set A against the diagnostic categories

Feature set A	Acriflavine		Fluorescein	
	Chi squared statistic	**p-value**	**Chi squared statistic**	**p-value**
cell crowding	36.92	<0.001	52.18	<0.001
cell size	54.18	<0.001	51.20	<0.001
nuclei crowding	51.81	<0.001	26.68	<0.001
nuclei size	56.84	<0.001	63.13	<0.001
fluorescent granules	11.33	0.07	18.40	0.005

Cell crowding, cell size, nuclei crowding, and nuclei size showed significant relationships with the diagnostic categories for the acriflavine images ($p < 0.001$), with nuclei size having the strongest association (chi-squared = 56.84). Fluorescent granules did not show a significant relationship ($p = 0.079$) in the acriflavine images and thus might not be strongly associated with the diagnostic categories (Table 5.2.).

Cell crowding, cell size, nuclei crowding, and nuclei size also showed strong relationships for the fluorescein images ($p < 0.001$), with nuclei size again showing the highest chi-squared statistic (63.13). Fluorescent granules were significantly different across categories in the fluorescein images ($p = 0.005$) in relation to the diagnostic categories but has the lowest chi-squared statistic (18.40), indicating a weaker association relative to the other features (Table 5.2.).

Nuclei size consistently showed the strongest association with the target variable across both conditions. Fluorescent granules had the weakest relative relationship overall, with no significance in the acriflavine images and marginal significance in the fluorescein images. Cell crowding, cell size, and nuclei crowding had moderate to strong associations in both contrast agent image sets (Table 5.2.).

5.4.2. Feature set B – Binary categorisation

The chi-squared test revealed that cell crowding, cell size, nuclei crowding, and nuclei size showed strong associations with the diagnostic category for the acriflavine images ($p < 0.001$), with nuclei size having the strongest association (chi-squared = 38.43). Fluorescent granules were significant but weaker ($p = 0.019$), indicating a less prominent role in acriflavine images (Table 5.3.).

Table 5.3. Chi-squared test results for feature set B against the diagnostic categories

Feature set B	Acriflavine		Fluorescein	
	Chi squared statistic	**p-value**	**Chi squared statistic**	**p-value**
cell crowding	24.15	<0.001	41.61	<0.001
cell size	22.55	<0.001	25.34	<0.001
nuclei crowding	34.96	<0.001	5.80	0.122
nuclei size	38.43	<0.001	4.45	0.217
fluorescent granules	9.85	0.019	13.27	0.004

Cell crowding and size maintained strong associations with the diagnostic category in the fluorescein images ($p < 0.001$), with cell crowding showing the strongest association (chi-squared = 41.61). Nuclei crowding and size did not show significant relationships in the fluorescein images ($p = 0.122$ and $p = 0.217$, respectively), suggesting their relevance depends on the contrast agent used. Fluorescent granules were significant in fluorescein images ($p = 0.004$) but remains the weakest among the significant features (Table 5.3.).

Cell crowding and size were consistently significant across both contrast agents in feature set B with moderate to strong associations. Nuclei crowding and size were significant only in the acriflavine images, highlighting their dependence on imaging contrast agent. Fluorescent granules while significant, showed a consistently weaker relationship compared to other features.

Overall, across both feature sets showed that qualitative analysis of acriflavine images consistently demonstrates strong associations between cell and nuclei characteristics (particularly nuclei size) and the diagnostic categories. While fluorescent granules showed limited diagnostic value, making this contrast agent particularly useful for assessing cellular and nuclear abnormalities in oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC). The qualitative analysis of fluorescein dataset showed strong associations for cell characteristics across feature sets but demonstrates inconsistency in the significance of nuclei features and weak but consistent associations for fluorescent granules. This dataset showed potential in evaluating cellular architecture while offering complementary information on nuclei and granular features in the diagnosis of OED and OSCC.

5.5. Machine learning diagnostic analysis of both contrast agent datasets

All micrographical features identified by the 2 oral medicine clinicians across the binary and multi-category feature sets (feature sets 'A' and 'B') were prepared to constitute the training and test datasets for the development of diagnostic prediction ML models. The 4 ML model types of logistic regression, support vector machines (SVM), random forest and XGBoost were developed and tested on both feature sets of acriflavine and fluorescein datasets.

5.5.1. Feature set A - multi-category

5.5.1.1. ML model performance on acriflavine feature set A

The test performance results of all 4 types of ML models on the acriflavine feature set A across the metrics of accuracy, sensitivity, specificity, precision and F1 score are depicted in Table 5.4.

Upon ranking all models based on test performance in all classes across all metrics the random forest and XGBoost models both had the best rank (Table 5.5.). However, upon comparison of the individual class results the random forest model was chosen as the best model due to its higher F1 score for high-risk samples (0.18 vs 0) (Table 5.4.).

Table 5.4. Performance results of ML models on the Acriflavine feature set A dataset

Class	Metric	Logistic regression	SVM	Random Forest	XGBoost
No dysplasia	Accuracy	0.53	0.53	0.60	0.57
	Sensitivity	0.84	0.84	0.80	0.84
	Specificity	0.31	0.31	0.46	0.37
	Precision	0.47	0.47	0.51	0.49
	F1 Score	0.60	0.60	0.63	0.62
Lichenoid	Accuracy	0.78	0.78	0.80	0.82
	Sensitivity	0.00	0.00	0.09	0.09
	Specificity	**0.96**	**0.96**	**0.96**	**0.98**
	Precision	0.00	0.00	0.33	0.50
	F1 Score	0.00	0.00	0.14	0.15
Low-risk	Accuracy	0.72	0.77	0.78	0.78
	Sensitivity	0.39	0.33	0.50	0.50
	Specificity	**0.86**	**0.95**	**0.90**	**0.90**
	Precision	0.54	0.75	0.69	0.69
	F1 Score	0.45	0.46	0.58	0.58
High-risk	Accuracy	**0.90**	**0.85**	**0.85**	**0.87**
	Sensitivity	0.00	0.17	0.17	0.00
	Specificity	**1.00**	**0.93**	**0.93**	**0.96**
	Precision	0.00	0.20	0.20	0.00
	F1 Score	0.00	0.18	0.18	0.00

Table 5.5. Ranking all ML models based on test performance metrics across all classes on the Acriflavine feature set A

Class	Metric	Logistic regression	SVM	Random Forest	XGBoost
No dysplasia	Accuracy	3	3	1	2
	Sensitivity	1	1	4	1
	Specificity	3	3	1	2
	Precision	3	3	1	2
	F1 Score	3	3	1	1
Lichenoid	Accuracy	3	3	2	1
	Sensitivity	3	3	1	1
	Specificity	2	2	2	1
	Precision	3	3	2	1
	F1 Score	3	3	2	1
Low-risk	Accuracy	4	3	1	1
	Sensitivity	3	4	1	1
	Specificity	4	1	2	2
	Precision	4	1	2	2
	F1 Score	4	3	1	1
High-risk	Accuracy	1	3	3	2
	Sensitivity	3	1	1	3
	Specificity	1	3	3	2
	Precision	3	1	1	3
	F1 Score	3	1	1	3
Aggregate rank		57	48	33	33
Final rank		4	3	1	1

The results of the best ranking Random Forest model are as follows:

No dysplasia:

Had the best F1 score across all classes (0.62). The model was moderately good at balancing the identification of no dysplasia while avoiding misclassification of other classes as no dysplasia (Table 5.4.). Had high sensitivity (0.8) but low specificity (0.46). Majority of no dysplasia images were correctly identified but all other classes were not clearly differentiated from 'no dysplasia' (Table 5.4.). A moderate precision (0.51) indicated the model got its positive 'no dysplasia' predictions correct about half the time (Table 5.4.).

Lichenoid:

The model had a poor performance with identifying lichenoid samples with an F1 score of 0.14 (Table 5.4.). A high specificity (0.96) indicates that the model was good at identifying every other class as not being 'lichenoid' (Table 5.4.).

Low-risk:

The model had a moderate overall performance with an F1 score of 0.58, indicating it can balance precision and recall but is not highly reliable at detection low-risk lesions (Table 5.4.). A high specificity (0.90) indicates the model's superior ability to correctly identify other classes as not being low-risk (Table 5.4.). However, that is balanced by the low sensitivity (0.50) indicating the model's inability to identifying true low-risk cases (Table 5.4.).

High-risk:

The model had a poor performance with identifying high-risk samples with an F1 score of 0.18 (Table 5.4.). A high specificity (0.93) indicates the model's superior ability to correctly identify other classes as not being high-risk (Table 5.4.).

This model demonstrated varying performance across diagnostic categories, with the best performance in identifying no dysplasia cases, moderate performance for low-risk cases, with exceptionally poor performance for lichenoid and high-risk cases. This suggests that while the model has some utility, it requires significant improvement, particularly for distinguishing between higher-risk category samples.

5.5.1.2. ML model performance on Fluorescein feature set A

The test performance results of all 4 types of ML models on the fluorescein feature set A across the metrics of accuracy, sensitivity, specificity, precision and F1 score are depicted in Table 5.6.

Table 5.6. Performance results of ML models on the Fluorescein feature set A

Class	**Metric**	**Logistic regression**	**SVM**	**Random Forest**	**XGBoost**
No dysplasia	Accuracy	0.45	0.38	0.55	0.48
	Sensitivity	0.70	0.57	0.61	0.61
	Specificity	0.30	0.27	0.51	0.41
	Precision	0.38	0.32	0.44	0.39
	F1 Score	0.49	0.41	0.51	0.47
Lichenoid	Accuracy	0.68	0.65	0.70	0.68
	Sensitivity	0.31	0.12	0.19	0.19
	Specificity	0.82	0.84	0.89	0.86
	Precision	0.38	0.22	0.37	0.33
	F1 Score	0.34	0.16	0.25	0.24
Low-risk	Accuracy	0.83	0.80	0.72	0.77
	Sensitivity	0.00	0.00	0.20	0.00
	Specificity	1.00	0.96	0.82	0.92
	Precision	0.00	0.00	0.18	0.00
	F1 Score	0.00	0.00	0.19	0.00
High-risk	Accuracy	0.80	0.80	0.83	0.80
	Sensitivity	0.18	0.36	0.45	0.45
	Specificity	0.94	0.90	0.92	0.88
	Precision	0.40	0.44	0.56	0.45
	F1 Score	0.25	0.40	0.50	0.45

Upon ranking all models based on test performance in all classes across all metrics the random forest model ranked the highest (Table 5.7.).

Table 5.7. Ranking all ML models based on test performance metrics across all classes on the Fluorescein feature set A

Class	**Metric**	**Logistic regression**	**SVM**	**Random Forest**	**XGBoost**
No dysplasia	Accuracy	3	4	1	2
	Sensitivity	1	4	2	2
	Specificity	3	4	1	2
	Precision	3	4	1	2
	F1 Score	2	4	1	3
Lichenoid	Accuracy	2	4	1	2
	Sensitivity	1	4	2	2
	Specificity	4	3	1	2
	Precision	1	4	2	3
	F1 Score	1	4	2	3
Low-risk	Accuracy	1	2	4	3
	Sensitivity	2	2	1	2
	Specificity	1	2	4	3
	Precision	2	2	1	2
	F1 Score	2	2	1	2
High-risk	Accuracy	2	2	1	2
	Sensitivity	4	3	1	1
	Specificity	1	3	2	4
	Precision	4	3	1	2
	F1 Score	4	3	1	2
Aggregate rank		44	63	31	46
Final rank		2	4	1	3

The results of the best ranking Random Forest model are as follows:

No dysplasia:

The moderate sensitivity (0.61) indicated the model identified just over half of the ‘no dysplasia’ cases correctly (Table 5.6.). The moderate specificity (0.51) indicated the model misclassified several cases from other classes as ‘no dysplasia’ (Table 5.6.). The model had an overall moderate performance in identifying ‘no dysplasia’ with an F1 score of 0.51 (Table 5.6.).

Lichenoid:

The model had low sensitivity (0.19) and high specificity (0.89) indicating it did not detect many ‘lichenoid’ cases and misclassified very few cases of other classes as ‘lichenoid’ (Table 5.6.). The poor F1 score (0.25) indicated the model is not reliable at predicted ‘lichenoid’ cases (Table 5.6.).

Low-risk:

The model had a similarly low sensitivity (0.2) for the ‘low-risk’ samples with a high specificity (0.89) indicating it did not detect many ‘low-risk’ cases and misclassified very few cases of other classes as ‘low-risk’ (Table 5.6.).

High-risk:

The model had low sensitivity (0.45) and high specificity (0.92) indicating it did not detect many ‘high-risk’ cases and misclassified very few cases of other classes as ‘high-risk’ (Table 5.6.). The model performed poorly at identifying ‘high-risk’ cases with an F1 score of 0.5 (Table 5.6.).

This model demonstrated varying performance across diagnostic categories, showing moderate ability to identify no dysplasia cases, poor performance for lichenoid, low-risk, and high-risk cases. This indicates that this model is unfit for the detection of OED and OSCC.

5.5.2. Feature set B – Binary categorisation

5.5.2.1. ML model performance on acriflavine feature set B

The test performance results of all 4 types of ML models on the acriflavine feature set B across the metrics of accuracy, sensitivity, specificity, precision and F1 score are depicted in Table 5.8.

Table 5.8. Performance results of ML models on the Acriflavine feature set B dataset

Class	**Metric**	**Logistic regression**	**SVM**	**Random Forest**	**XGBoost**
No dysplasia	Accuracy	0.52	0.53	0.53	0.53
	Sensitivity	0.8	0.8	0.8	0.8
	Specificity	0.31	0.34	0.34	0.34
	Precision	0.45	0.47	0.47	0.47
	F1 Score	0.58	0.59	0.59	0.59
Lichenoid	Accuracy	0.78	0.78	0.78	0.78
	Sensitivity	0	0	0	0
	Specificity	0.96	0.96	0.96	0.96
	Precision	0	0	0	0
	F1 Score	0	0	0	0
Low-risk	Accuracy	0.73	0.75	0.77	0.77
	Sensitivity	0.39	0.44	0.44	0.44
	Specificity	0.88	0.88	0.9	0.9
	Precision	0.58	0.62	0.67	0.67
	F1 Score	0.47	0.52	0.53	0.53
High-risk	Accuracy	0.87	0.9	0.88	0.88
	Sensitivity	0	0.17	0.17	0.17
	Specificity	0.96	0.98	0.96	0.96
	Precision	0	0.5	0.33	0.33
	F1 Score	0	0.25	0.22	0.22

Table 5.9. Ranking all ML models based on test performance metrics across all classes on the Acriflavine feature set B

Class	**Metric**	**Logistic regression**	**SVM**	**Random Forest**	**XGBoost**
No dysplasia	Accuracy	4	1	1	1
	Sensitivity	1	1	1	1
	Specificity	4	1	1	1
	Precision	4	1	1	1
	F1 Score	4	1	1	1
Lichenoid	Accuracy	1	1	1	1
	Sensitivity	1	1	1	1
	Specificity	1	1	1	1
	Precision	1	1	1	1
	F1 Score	1	1	1	1
Low-risk	Accuracy	4	3	1	1
	Sensitivity	4	1	1	1
	Specificity	3	3	1	1
	Precision	4	3	1	1
	F1 Score	4	3	1	1
High-risk	Accuracy	4	1	2	2
	Sensitivity	4	1	1	1
	Specificity	2	1	2	2
	Precision	4	1	2	2
	F1 Score	4	1	2	2
Aggregate rank		59	28	24	24
Final rank		4	3	1	1

Upon ranking all models based on test performance in all classes across all metrics the Random Forest and XGBoost models ranked the highest (Table 5.9.). The performance of both models across all metrics and classes were identical (Table 5.8.). Their results are interpreted together.

The results of the random forest and XGBoost models on the acriflavine feature set B are as follows:

No dysplasia:

The models had moderate performance at detecting No dysplasia cases with a precision of 0.47 and F1 score of 0.59 (Table 5.8.). High sensitivity (0.80) indicates the model was good at identifying no dysplasia cases (Table 5.8.). However, poor specificity (0.34) indicates the model incorrectly classified several other class samples as 'no dysplasia' (Table 5.8.).

Lichenoid:

The model did not classify any 'lichenoid' case correctly with a sensitivity, precision, and F1 score of 0 (Table 5.8.).

Low-risk:

The model had a moderate performance at identifying 'Low-risk' cases with a precision of 0.67, sensitivity of 0.44, and F1 score of 0. (Table 5.8.). A high specificity (0.90) indicated the model performed well at identifying non 'Low-risk' cases as not being 'Low-risk' (Table 5.8.).

High-risk:

The model performed poorly at detecting 'High-risk' cases with a sensitivity of 0.17, and F1 score of 0.22 (Table 5.8.). It had a high specificity (0.96) which indicates the model misclassified very few other cases from another class into 'High-risk' (Table 5.8.).

The model shows moderate performance in identifying No dysplasia and Low-risk cases and poor performance on High-risk cases but completely fails to correctly classify Lichenoid. This indicates that this model is unfit for the detection of OED and OSCC.

5.5.2.2. ML model performance on fluorescein feature set B

The test performance results of all 4 types of ML models on the fluorescein feature set B across the metrics of accuracy, sensitivity, specificity, precision and F1 score are depicted in Table 5.10.

Table 5.10. Performance results of ML models on the Fluorescein feature set B

Class	Metric	Logistic regression	SVM	Random Forest	XGBoost
No dysplasia	Accuracy	0.47	0.30	0.30	0.30
	Sensitivity	0.52	0.57	0.57	0.57
	Specificity	0.43	0.14	0.14	0.14
	Precision	0.36	0.29	0.29	0.29
	F1 Score	0.43	0.38	0.38	0.38
Lichenoid	Accuracy	0.57	0.70	0.70	0.70
	Sensitivity	0.31	0.00	0.00	0.00
	Specificity	0.66	0.95	0.95	0.95
	Precision	0.25	0.00	0.00	0.00
	F1 Score	0.28	0.00	0.00	0.00
Low-risk	Accuracy	0.83	0.77	0.77	0.77
	Sensitivity	0.00	0.00	0.00	0.00
	Specificity	1.00	0.92	0.92	0.92
	Precision	0.00	0.00	0.00	0.00
	F1 Score	0.00	0.00	0.00	0.00
High-risk	Accuracy	0.80	0.80	0.80	0.80
	Sensitivity	0.27	0.36	0.36	0.36
	Specificity	0.92	0.90	0.90	0.90
	Precision	0.43	0.44	0.44	0.44
	F1 Score	0.33	0.40	0.40	0.40

The SVM, RF, and XGBoost models all had the exact same predictions and therefore identical test results, and they performed better than the logistic regression model (Table 5.11.).

Table 5.11. Ranking all ML models based on test performance metrics across all classes on the Fluorescein feature set B

Class	**Metric**	**Logistic regression**	**SVM**	**Random Forest**	**XGBoost**
No dysplasia	Accuracy	1	2	2	2
	Sensitivity	4	1	1	1
	Specificity	1	2	2	2
	Precision	1	2	2	2
	F1 Score	1	2	2	2
Lichenoid	Accuracy	4	1	1	1
	Sensitivity	1	2	2	2
	Specificity	4	1	1	1
	Precision	1	2	2	2
	F1 Score	1	2	2	2
Low-risk	Accuracy	1	2	2	2
	Sensitivity	1	1	1	1
	Specificity	1	2	2	2
	Precision	1	1	1	1
	F1 Score	1	1	1	1
High-risk	Accuracy	1	1	1	1
	Sensitivity	4	1	1	1
	Specificity	1	2	2	2
	Precision	4	1	1	1
	F1 Score	4	1	1	1
Aggregate rank		38	30	30	30
Final rank		4	1	1	1

The identical results of the SVM, random forest and XGBoost models are as follows:

No dysplasia:

The models had a poor performance at identifying ‘no dysplasia’ cases with a moderate sensitivity (0.57), and a poor specificity (0.14), precision (0.29) and F1 score (0.38) (Table 5.10.).

Lichenoid:

The model did not identify a single ‘Lichenoid’ case with a sensitivity, precision, and F1 score of 0 (Table 5.10.).

Low-risk:

The model did not identify a single ‘Low-risk’ case correctly with a sensitivity, precision, and F1 score of 0 (Table 5.10.).

High-risk:

The model correctly identified less than half of the ‘High-risk’ cases with a sensitivity of 0.36 and less than half of its ‘High-risk’ predictions were correct with a precision of 0.44 (Table 5.10.). A high specificity (0.90) indicated the model performed well at not misclassifying other classes as ‘High-risk’ (Table 5.10.). Overall, with an F1 score of 0.4 the model was not reliable at predicting ‘High-risk’ cases (Table 5.10.).

This model shows moderate to poor performance in identifying No dysplasia cases, and High-risk cases, and it completely failed to detect Lichenoid and Low-risk cases. This model was severely limited in accurately distinguishing between OED grades and is unfit for diagnostic triage.

5.6. The best ML models across all feature sets

The overall performance across all models was moderate to poor with respect to the 4 diagnostic classes (Figure 5.6.).

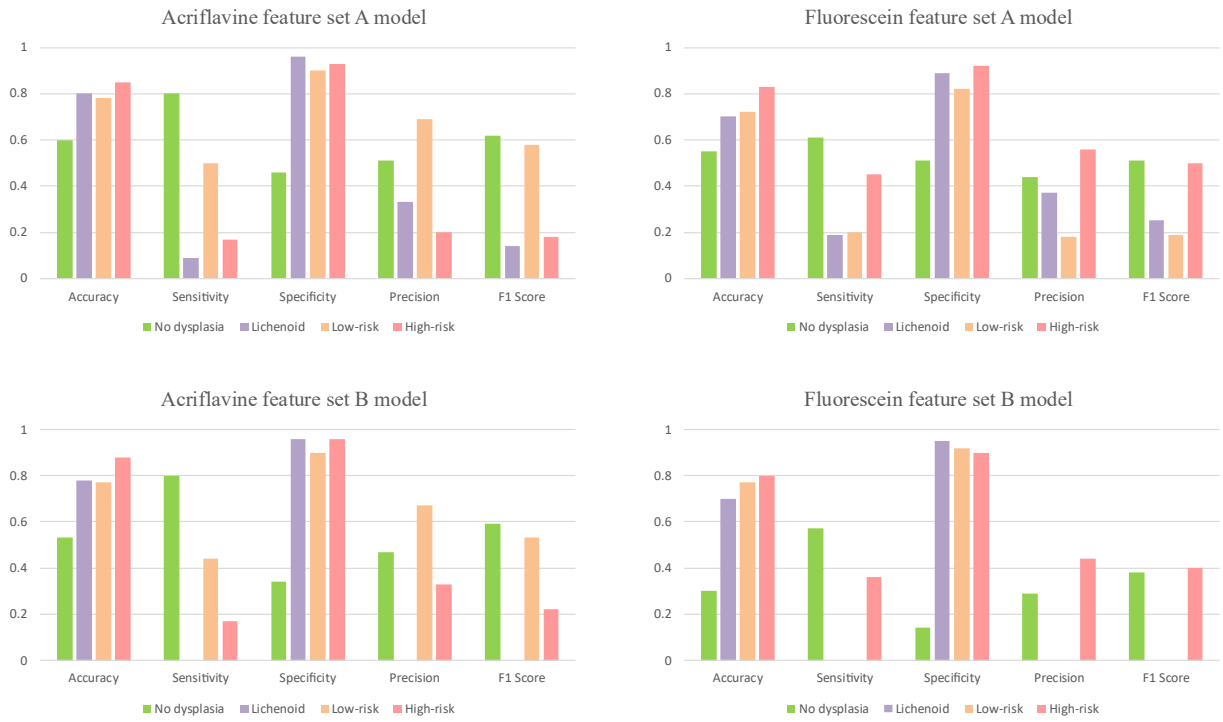


Figure 5.6. Bar graphs of best model test performance for each feature set

The overall performance of the best performing ML models in each feature set and contrast agent can be summarised as follows (Figure 5.6.):

No dysplasia:

The models appear to be good at detecting positive cases of no dysplasia (high sensitivity) while also often incorrectly labelling other classes as 'no dysplasia' (low precision and specificity) (Figure 5.6.).

Lichenoid:

The fluorescein feature set models failed to predict a single Lichenoid case correctly (Figure 5.6.). The acriflavine feature set models only showed high specificity but poor sensitivity, precision and F1 score (Figure 5.6.).

Low-risk:

All models had moderate to poor sensitivity and precision indicating about half of the ‘Low-risk’ sample correctly (Figure 5.6.).

High-risk:

All feature set models had low to moderate sensitivity and precision where they identified less than half of the actual high-risk cases correctly (Figure 5.6.).

Table 5.12. Ranking the best models from each acriflavine and fluorescein feature set

Class	Metric	Acriflavine feature set A model	Fluorescein feature set A model	Acriflavine feature set B model	Fluorescein feature set B model
No dysplasia	Accuracy	1	2	3	4
	Sensitivity	1	3	1	4
	Specificity	2	1	3	4
	Precision	1	3	2	4
	F1 Score	1	3	2	4
Lichenoid	Accuracy	1	3	2	3
	Sensitivity	2	1	3	3
	Specificity	1	4	1	3
	Precision	2	1	3	3
	F1 Score	2	1	3	3
Low-risk	Accuracy	1	4	2	2
	Sensitivity	1	3	2	4
	Specificity	2	4	2	1
	Precision	1	3	2	4
	F1 Score	1	3	2	4
High-risk	Accuracy	2	3	1	4
	Sensitivity	3	1	3	2
	Specificity	2	3	1	4
	Precision	4	1	3	2
	F1 Score	4	1	3	2
Aggregate rank		35	48	44	64
Final rank		1	3	2	4

After ranking all the best models for each feature set, the acriflavine feature set A model (random forest) performed the best (Table 5.12.).

The results of the overall best performing model in diagnostic classification of human identified features on in vivo confocal micrographs are depicted in Table 5.13. and Table 5.14.

Table 5.13. Test confusion matrix for best performing model (Acriflavine feature set A - random forest) for diagnostic classification of human identified features

		Predicted			
		No dysplasia	**Lichenoid**	**Low-risk**	**High-risk**
Actual	**No dysplasia**	20	2	2	1
	Lichenoid	8	1	1	1
	Low-risk	7	0	9	2
	High-risk	4	0	1	1

No dysplasia:

For the no dysplasia group (n=25), the model achieved a sensitivity of 0.80 but relatively low specificity (0.46), leading to frequent misclassification of other conditions as non-dysplastic. Precision was modest (0.51), yielding an F1 score of 0.63 (Table 5.13. & Table 5.14.).

Lichenoid:

Performance in the lichenoid class (n=11) was poor, with sensitivity of only 0.09 and an F1 score of 0.14. Although specificity was high (0.96), the model failed to identify most true lichenoid cases, reflecting strong under-detection (Table 5.13. & Table 5.14.).

Low-risk:

In the low-risk dysplasia group (n=18), sensitivity reached 0.50 with specificity of 0.90, precision of 0.69, and the highest F1 score among the dysplastic categories (0.58). This suggests the model was better at distinguishing low-risk dysplasia relative to other pathological groups (Table 5.13. & Table 5.14.).

High-risk:

The high-risk dysplasia class (n=6) showed limited detectability, with sensitivity of 0.17, specificity of 0.93, and F1 score of 0.18, indicating both low recall and low precision in recognising high-risk lesions (Table 5.13. & Table 5.14.).

Table 5.14. Performance metrics for best performing model (Acriflavine feature set A - random forest) for diagnostic classification of human identified features

Class	Sensitivity	Specificity	Precision	F1 score
No dysplasia	0.80	0.46	0.51	0.63
Lichenoid	0.09	0.96	0.33	0.14
Low-risk	0.50	0.90	0.69	0.58
High-risk	0.17	0.93	0.20	0.18

This model had a moderate overall AUROC performance with the highest AUC of 0.74 for ‘Low-risk’ and ‘High-risk’ categories, with ‘No dysplasia’ trailing at 0.71 and ‘Lichenoid’ with the least at 0.68 (Figure 5.7.).

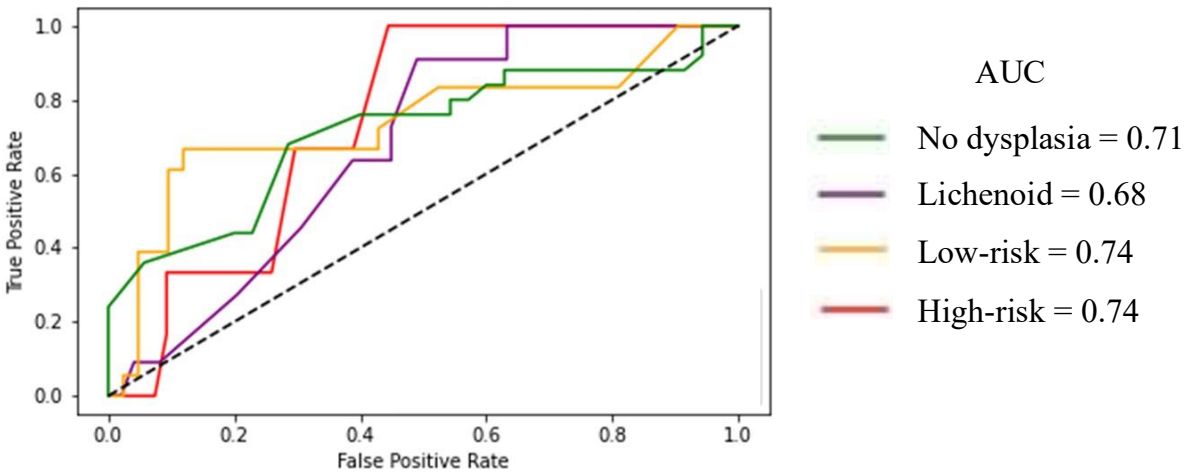


Figure 5.7. AUROC of the best performing machine learning model on the qualitative human identified features (acriflavine feature set A model)

Table 5.15. Best model (acriflavine feature set A - random forest) performance in identifying OED and OSCC using human-identified features

	Cases correctly identified (sensitivity)	**Other categories wrongly predicted as this category** (1-specificity)	**Predictions of this category that were incorrect** (1 - precision)	**Cases misclassified** (1-sensitivity)
Low-grade OED	50%	9.52%	30.77%	50%
High-grade OED & OSCC	16.67%	7.41%	80%	83.33%

From a clinical decision-making perspective of recommending urgent and non-urgent biopsies, this acriflavine feature set A model identified only 50% of low-grade OED cases and only 16.67% of high-grade OED and OSCC cases (Table 5.15.).

5.7. Discussion

This chapter describes the first diagnostic study in this dissertation. Microscopic oral epithelial qualitative features reported by previous confocal microscopy studies in the systematic review (Chapter 2) were recorded as inputs for diagnostic prediction machine learning (ML) models (Ramani et al., 2023). The present study involved using qualitative human-identified features such as cell size homogeneity, cell crowding, nucleus size homogeneity, nuclear crowding, and presence of fluorescence granules to train ML models for diagnostic classification of OED and OSCC.

There were some statistically significant associations with the diagnostic categories and nucleus and cellular features, including OED and OSCC. The chi-squared tests showed significant (p value <0.05) differences between diagnostic categories across both contrast agents and feature sets. All findings the cellular and nuclear features showed significant differences except for nuclear features in fluorescein feature set B and fluorescent granules for acriflavine feature set A. Other than these exceptions there were no differences between the two feature sets.

The significantly differences in features observed across acriflavine and fluorescein datasets could be due to the differences in their mechanisms of staining oral epithelial tissue. Acriflavine highlights epithelial cell nuclei primarily by intercalating with Deoxyribonucleic acid (DNA), providing direct visualization of nuclear morphology within the cells (Piorecka et al., 2022). When using fluorescein to stain cells, its charged ends become attracted to the hydrophilic ends in the cell membrane to form a strong electrostatic bond. This allows fluorescein to be a pan-cytoarchitectural contrast agent that highlights a range of oral epithelial structures (Ragazzi et al., 2014).

These features depict a direct link between known microscopic markers of oral carcinogenesis as previously described (E. Odell, O. Kujan, S. Warnakulasuriya, & P. Sloan, 2021a) and the diagnosis of OED & OSCC. Altered proliferation of cells, indicated by increased mitotic activity and loss of epithelial organisation in the form of irregular stratification, agrees with the cell and nuclear crowding observations in the present study. Cytological atypia in the form of abnormal dysplasia variation in cell and nuclei size along with a skewed nucleus - cytoplasmic ratio are standard histopathology markers for carcinogenic processes (Odell et al., 2021a). This is represented by the variation in cell and nuclei size homogeneity that was observed in the confocal micrographs in this study.

The variability of these dysplastic changes across lesions of the same diagnostic category are also a feature of oral carcinogenesis. High-risk lesions, such as high-grade OED and OSCC, displayed in this study high levels of cell and nuclear crowding along with a heterogeneity of cell and nuclei sizes in alignment with previously observed signs of oral carcinogenesis in histopathology (E. Odell, O. Kujan, S. Warnakulasuriya, & P. J. O. d. Sloan, 2021b). Qualitative confocal microscopy studies from the systematic review in Chapter 2 on oral cancer detection had also observed similar cytologic architectural features in high-risk lesions (Ramani et al., 2023).

Machine learning (ML) models using these features as inputs were expected to capture these patterns and classify previously unseen images with the observed cellular and nuclear features into the defined OED and OSCC diagnostic categories. For robustness 4 different ML model types; logistic regression (LR), support vector machines (SVM), random forest (RF), and XGBoost (XGB); were used for this analysis to cover a range of ML approaches from linear to non-linear decision boundaries, simple to complex architectures, and different ensemble techniques.

All ML models performed poorly across both the binary and multi-category feature sets for both contrast agent datasets. For cases of no dysplasia, the models exhibit high sensitivity in detecting positive cases, but they often incorrectly label other classes as 'no dysplasia', resulting in low precision and specificity. The performance for lichenoid cases is particularly poor, with the fluorescein feature set models failing to correctly predict any lichenoid cases, while the acriflavine feature set models show high specificity but poor sensitivity, precision, and F1 score. For low-risk cases, all models display moderate to poor sensitivity and precision, correctly identifying about half of the 'low-risk' samples. Similarly, in high-risk cases, all feature set models show low to moderate sensitivity and precision, correctly identifying less than half of the actual high-risk cases. Overall, the models were inconsistent and unreliable for predicting the lichenoid, OED and OSCC cases, miscategorising a majority of these as ‘no dysplasia’. The acriflavine feature set A ML model (random forest) performed the best. However, it has very limited clinical utility since it identified only 50% of low-grade OED cases and 16.67% high-grade OED and OSCC cases.

Human annotations in medical images can often vary between experts leading to potentially inconsistent labels (Yang et al., 2023). Any disagreement on the presence of extent of findings in an image, such as the cellular and nuclear features in the confocal microscopy images in the present study, could introduce noise in the training data that in turn impacts ML model performance (Koçak et al., 2025). The novelty of fluorescence in vivo

confocal microscopy and the lack of specific standards on identification of features of the oral epithelium make annotation quite challenging. Additionally, the process of labelling images in this study was inherently subjective. Different experts could have varying thresholds on what constitutes an abnormality. This is called annotator bias, and it can be influenced by personal biases and level of experience in a given field (Koçak et al., 2025). Annotator bias can skew labels and thus the model's learning process. Any of these factors could have impacted the training data for the ML models in this study and thus could have led to their poor overall performance.

Human annotation of medical images is also labour-intensive, time-consuming, and costly (Zhang & Qie, 2023). This especially applies to the images in the present study that involved careful screening of microscopy images for small changes in cellular and nuclear elements of the oral epithelium. Due to these constraints of human annotators a small dataset of randomly selected 600 images were included that potentially limited ML model performance as these algorithms require large amounts of data for learning (Zhang & Qie, 2023). Additionally, a limitation of this study is that the random selection of test images resulted in an uneven distribution of cases across diagnostic categories. Although random allocation avoids systematic bias, in a relatively small dataset this can inadvertently under-represent certain classes, thereby affecting the reliability of class-specific performance estimates. Because the test set was sampled in proportion to the database distribution, rarer categories such as high-risk and lichenoid were under-represented, making their performance metrics less reliable despite the overall sampling being unbiased

An approach for improving this study would involve additional expert annotators of varying levels of experience with confocal microscopy to label these images in an ensemble to reduce annotator bias. While the range of ML models tested represented a spectrum of approaches, there are other models, such as Bayesian approaches, and instance based neighbours classifiers that could have a different performance (Abdullah et al., 2022). This constantly increasing list of ML models were not developed in this study due to limitation of time, resources, and compute power.

The integration of machine learning with human expertise in chairside clinical diagnostic systems presents a promising approach for enhancing the early detection and triage of OED and OSCC. This model combines the consistent pattern recognition and rapid data processing of machine learning with the contextual judgment and clinical experience of dental practitioners. Such hybrid systems promise numerous benefits, including reduced variability in assessments, and support for informed decision-making. Importantly, it would

maintain human oversight in critical decisions, ensuring accountability, and patient safety. The preservation of clinician autonomy in AI-assisted diagnoses might also help to foster patient trust. Ultimately, this collaborative model not only improves diagnostic outcomes but also has the potential to revolutionise oral cancer detection and management, marking a significant advancement in dental healthcare.

However current results indicates that these ML models assessing human identified qualitative features are not fit for clinical use. Since the link between oral epithelial features on fluorescence in vivo confocal microscopy and OED & OSCC was identified, future work might involve developing and validating a quantitative analysis methodology with machine augmented systematic annotation of features.

6. MACHINE LEARNING FOR QUANTITATIVE FEATURE EXTRACTION

6.1. Introduction

Oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) represent significant global health challenges, with early detection playing a crucial role in improving patient outcomes. Traditional diagnostic methods often rely on invasive biopsies and subjective histopathological assessments that can be time-consuming and prone to inter-observer variability. The WHO classification of head and neck tumours lists nuclear shape and nuclear size on histopathology microscopy sections as signs of OED (Reibel et al., 2017a).

A review by Nag & Das (2018) outlines studies that have used image analysis of histopathology sections for oral cancer detection assessing the presence of abnormal nuclear size (anisonucleosis), hyperchromatism, and irregularities in nuclear shape (nuclear pleomorphism) (Nag & Das, 2018). Morphometric analysis of nuclei of histopathology images using image analysis software and measurements of hand-labelled nuclei have shown that normal nuclei are round with regular membrane outlines, whereas they increasingly become oval and irregular as the grade of dysplasia increases (Gadiwan et al., 2014). These approaches involve biopsies that are invasive and require several steps from sample collection to processing in a laboratory, imaging, and analysis, a long process subject to potential handling and processing errors.

There is a pressing need for more efficient, objective, and non-invasive diagnostic techniques to address the limitations of traditional methods. Confocal laser endomicroscopy (CLE) has emerged as one such promising non-invasive imaging modality for visualizing cellular structures in real-time. This technique allows for high-resolution imaging of the oral mucosa, enabling observation of cellular and tissue architecture without tissue removal (Yap et al., 2023). This imaging technique is augmented with the use of fluorophores to enhance image quality.

Two classic dyes, acriflavine and fluorescein, have long been used as topical contrast agents to visualize cells and tissues under fluorescent illumination. Acriflavine, an acridine-derived dye, preferentially stains nuclei with a yellow-green fluorescence, whereas fluorescein, a xanthene dye, diffuses through tissues and emits bright green fluorescence (Piorecka et al., 2022; Robertson et al., 2013). Both compounds have rich histories in science and medicine and remain relevant in modern imaging techniques (Piorecka et al., 2022; Robertson et al., 2013). However, the interpretation of these images requires expertise and can still be subject to variability.

A few studies have explored the use of quantitative image analysis techniques with CLE for OED and OSCC detection (Ramani et al., 2023). Carlson et al.

(2007) analysed the fluorescence labelling intensity of epithelial structures using MATLAB, achieving a sensitivity and specificity of 1 and 0.86, respectively (Carlson, Gillenwater, Williams, El-Naggar, & Richards-Kortum, 2007b). Dittberner et al. (2016) developed an algorithm to identify cell size patterns between normal and OSCC tissue images, with a sensitivity and specificity of 0.72 and 0.85, respectively (Dittberner et al., 2016a). Other studies have examined cell density, nucleus density, connective tissue blood vessel diameter, and autofluorescence intensity measurements (Anuthama et al., 2010; Lupu et al., 2020; Shinohara et al., 2020).

Recent advancements in machine learning and image analysis offer the potential to enhance the diagnostic capabilities of in vivo confocal microscopy. One such approach involves segmentation of objects of interest (such as nuclei) through the process of partitioning an image into meaningful regions and distinguishing individual objects from the background in microscopy images to achieve precise boundary detection. Traditional image processing techniques, such as thresholding and watershed algorithms, often struggle with the challenges posed by microscopy images, including variability in nuclei shapes, low contrast, overlapping structures, and noise (Caicedo et al., 2019). To overcome these limitations, deep learning-based segmentation models have emerged as powerful tools for accurate and efficient nuclei segmentation.

Deep learning models for nuclei segmentation can be broadly categorized into semantic segmentation and instance segmentation. Semantic segmentation models, such as U-Net, classify each pixel in an image as either belonging to the foreground or background. Instance segmentation models not only classify pixels but also assign unique labels to individual objects, making them more suitable for analysing clustered or overlapping nuclei structures. A competition for nuclei segmentation called Data Science Bowl (2018) involved 3891 teams worldwide developing algorithmic solutions. The winning models included the U-Net architecture (F1 score = 0.89) accurately identifying nuclei in fluorescence microscopy images (Caicedo et al., 2019). StarDist 2D is a specialised neural network model inspired from this U-Net architecture that identifies star-convex polygons by predicting a set of radial distances from the centres of objects to their boundaries, allowing for efficient and shape-aware segmentation of nuclei with irregular contours (Schmidt et al., 2018). Cellpose is another modern deep learning segmentation method based on the U-Net architecture. It is a generalist algorithm that can precisely segment cells of a wide range of images without requiring parameter adjustments (Stringer et al., 2021).

By leveraging such advanced image analysis techniques and machine learning algorithms, the present study attempts to use nuclei measurements to create a

more objective and efficient method for diagnosing oral mucosal lesions. The focus was on the extraction and quantitative analysis of epithelial cell nuclei features from fluorescence in vivo confocal microscopy images of the oral mucosa using segmentation. This approach has the potential to improve the accuracy and speed of OED and OSCC diagnosis, ultimately leading to better patient outcomes and more effective treatment strategies.

Aim: To develop and evaluate machine learning models capable of classifying in vivo confocal microscopy images into lichenoid lesions, low- and high-grade oral epithelial dysplasia and oral squamous cell carcinoma based on quantitative features extracted from epithelial cell nuclei.

6.2. Methods

6.2.1. Cohort and imaging

In vivo confocal micrographs analysed in this study were captured prospectively in patients undergoing standard histopathological examination for oral mucosal abnormalities attending the Oral Medicine Department of the Royal Dental Hospital of Melbourne. Following imaging, oral mucosal lesions were biopsied and diagnosed using standard care histopathological assessment by a qualified pathologist.

Images were derived from lesions categorised in the following diagnostic classes (Table 3.2. from Chapter 3):

1. **No dysplasia** - All lesions that do not show any signs of dysplasia such as amalgam tattoo, chronic inflammation, denture hyperplasia, fibroepithelial polyp, focal papillomatosis, hyperplasia, hyperkeratosis, squamous papilloma, and verrucous xanthoma

2. **Lichenoid** - Oral lichen planus and lichenoid inflammation

3. **Low-risk** - Low-grade oral epithelial dysplasia (OED) with atypia and verrucous hyperplasia

4. **High-risk** - All high-grade OED lesions and oral squamous cell carcinoma (OSCC)

The in vivo CLE imaging was conducted using the InVivage® (Optiscan Imaging, Victoria Australia), a point scanning, fluorescence confocal endomicroscope with hand- and foot-called control (Protocol 1 described in Chapter 3). Acriflavine (0.1%) and fluorescein (0.1%) were the topical contrast agents used in this study. Images were captured from mucosae of the tongue, buccal mucosa, gingiva & vestibule, floor of mouth, hard palate, and soft palate.

6.2.2. Dataset

A set of randomly selected 600 images equally divided between the acriflavine (n=300) and fluorescein (n=300) image sets were taken from the overall dataset that was filtered for quality using the trained quality filtering convolutional neural network (PyTorch QMR) as described in Chapter 4.

Images were measured for size, shape, pixel intensity and spatial relationships of epithelial cell nuclei.

It was divided into 3 experiments with the following objectives:

1) To train and compare test performance of two different nucleus segmentation models on fluorescence in vivo confocal microscopy images of the oral mucosa epithelium.

2) To develop a segmentation model that can accurately identify nuclei on fluorescence in vivo confocal microscopy images of the oral mucosa epithelium.

3) To use nuclei measurements from nuclei identified by the developed segmentation model on fluorescence in vivo confocal microscopy images of the oral mucosa epithelium with machine learning models for classification of oral epithelial dysplasia and oral squamous cell carcinoma.

6.2.3. Experiment 1: Segmentation model comparison

Two existing segmentation models Cellpose 2D and StarDist 2D were compared for demarcating nuclei in fluorescence in vivo confocal micrographs by designing a bespoke annotation workflow.

This process involved a human-in-the-loop refinement process where the images were initially annotated using the Cellpose software with the cyto3 generalist algorithm (pre-trained) to identify nuclei. This automated annotation process to establish the target labels for segmentation training was complemented by a manual annotation step conducted by a researcher (R.R.), who added 20-30 additional annotations per image and corrected any errors in the initial automated annotations.

A total of 40 in vivo confocal microscopy images containing 10 images representing each of the classes: no dysplasia, lichenoid, low-risk and high-risk were selected for this initial experiment. This hybrid approach of using a pre-trained segmentation model in Cellpose 2D with human intervention was used to annotate all 32 training images (80%) and 8 test images (20%). The ZeroCostDL4Mic Google Colaboratory platform was used to train and validate Cellpose 2D and StarDist 2D models. Model performance was evaluated using the following metrics:

1) **Intersection over union (IoU) -** The overlap between the initial automated predicted objects and the subsequent human annotated objects with the value of '1' representing a perfect overlap (Figure 3.11. from Chapter 3).

2) **Precision** - the proportion of correct positive predictions
3) **Recall** - the total instances of positive samples that were correctly predicted
4) **F1 scores** - The harmonic mean of precision and recall

Training and validation loss, which are measures of the errors made by the segmentation models during the training phase, were plotted. Statistical analysis to compare performance of the Cellpose 2D and StarDist 2D models was undertaken using the paired sample t-test. This study determined which of the two models were superior on the nuclei segmentation task.

6.2.4. Experiment 2: Segmentation model development

This experiment involved using the bespoke human-in-the-loop annotation workflow developed in experiment #1 (Section 3.5. in Chapter 3) to label 400 acriflavine and fluorescein in vivo confocal microscopy images from 14 participants for training and testing using the best nucleus segmentation model from experiment #1 above. The dataset was divided into 360 training (90%) and 40 testing images (10%).

The training and evaluation of the chosen segmentation model was conducted using the ZeroCostDL4Mic platform that is designed to make deep learning more accessible to the microscopy community.

Model performance was evaluated using the following metrics:

1. **Accuracy:** The proportion of correct predictions out of total predictions
2. **F1 score:** The harmonic mean of precision and recall, offering a balanced measure of performance
3. **Mean true score (MTS):** Quantifies how well the overall predicted segments overlap the ground truth segments by taking the mean of the IoUs across an entire image
4. **Mean matched score (MMS):** Measures the average IoU score of predicted objects that have a corresponding matched ground truth object (IoU threshold >0.5)
5. **Panoptic quality (PQ):** A comprehensive metric (similar to F1 score) that evaluates both the detection and segmentation quality of the mode

6.2.5. Experiment 3: Machine learning of extracted nuclear features for diagnostic triage

This experiment aimed to develop, evaluate, and compare ML models for diagnostic classification of feature extracted data of epithelial cell nuclei segmented by the model developed in the previous experiment from fluorescence human in vivo captured confocal endomicroscopy images (Figure 6.1.). The developed model was chosen to segment nuclei in a dataset composed of both acriflavine and fluorescein augmented in vivo confocal micrographs across the 4 diagnostic triage categories (Table 3.2. from Chapter 3).

The metrics measured of the nuclei identified in both the acriflavine and fluorescein datasets were (Table 3.8. from Chapter 3):

1. **Area:** Region of selection in square pixels
2. **Mean pixel intensity:** Average pixel brightness (intensity) value (range = 0-255) within the selection
3. **Pixel intensity standard deviation:** Standard deviation of pixel brightness/pixel intensity values
4. **Circularity:** A value of 1.0 indicates a perfect circle. As the value approaches 0.0, it indicates an increasingly elongated shape. *Formula:* $4\pi * \frac{area}{perimeter^2}$
5. **Aspect ratio (AR):** Ratio of length of major axis to length of minor axis
6. **Integrated density:** Raw integrated density is the sum of the values of the pixels in the image or selection

Summary statistics of mean, median, and standard deviation were calculated for all measurements across all diagnostic categories using Fiji – ImageJ image analysis software (Schindelin et al., 2015). The one-way ANOVA was undertaken to determine whether the mean measurement values for all the images differed significantly across the diagnostic classes. The Spearman's correlation was undertaken to assess the strength and direction of the relationships between pairs of measurements.

This experiment employs three main approaches for machine learning analysis:

1. **Approach 1:** involved direct classification of measurement means that were z-score standardized.
2. **Approach 2:** utilized clustering and feature selection, where k-means clustering was applied to group nuclei based on their measurements, and the resulting cluster proportions per image are used as features for classification.
3. **Approach 3:** used spatial distribution of nuclei through Euclidean distance-based features, measuring distances between nuclei as input for the ML models

Four types of machine learning models are developed and compared (Table 3.5. from Chapter 3):

1. **Logistic Regression (LR):** A simple yet powerful linear model that assumes a linear decision boundary and is effective for well-separated data (Hosmer Jr, Lemeshow, & Sturdivant, 2013)
2. **Support Vector Machine (SVM):** A model that constructs hyperplanes to separate classes in a high-dimensional space, effective for non-linear separations when paired with kernel functions (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998)
3. **Random Forest (RF):** An ensemble model that combines multiple decision trees to improve prediction accuracy and reduce overfitting (Breiman, 2001)
4. **XGBoost (XGB):** A ensemble model known for its high efficiency and performance, particularly in structured data tasks (Chen & Guestrin, 2016)

These ML models represent a spectrum of approaches to classification from simple to complex and linear to decision tree-based predictions. Each model underwent extensive hyperparameter tuning using grid search methods to optimize performance. The dataset is split into 80% training and 20% testing sets, with stratified 5-fold cross-validation during training to ensure robust model evaluation. Feature selection and dimensionality reduction techniques were employed, particularly in the clustering approach (as described in Section 3.5.3. in Chapter 3). This experiment aimed to determine which combination of feature extraction, data preprocessing, and ML model across in vivo fluorescence CLE images across both acriflavine, and fluorescein datasets yields the most accurate and reliable diagnostic predictions for OED and OSCC.

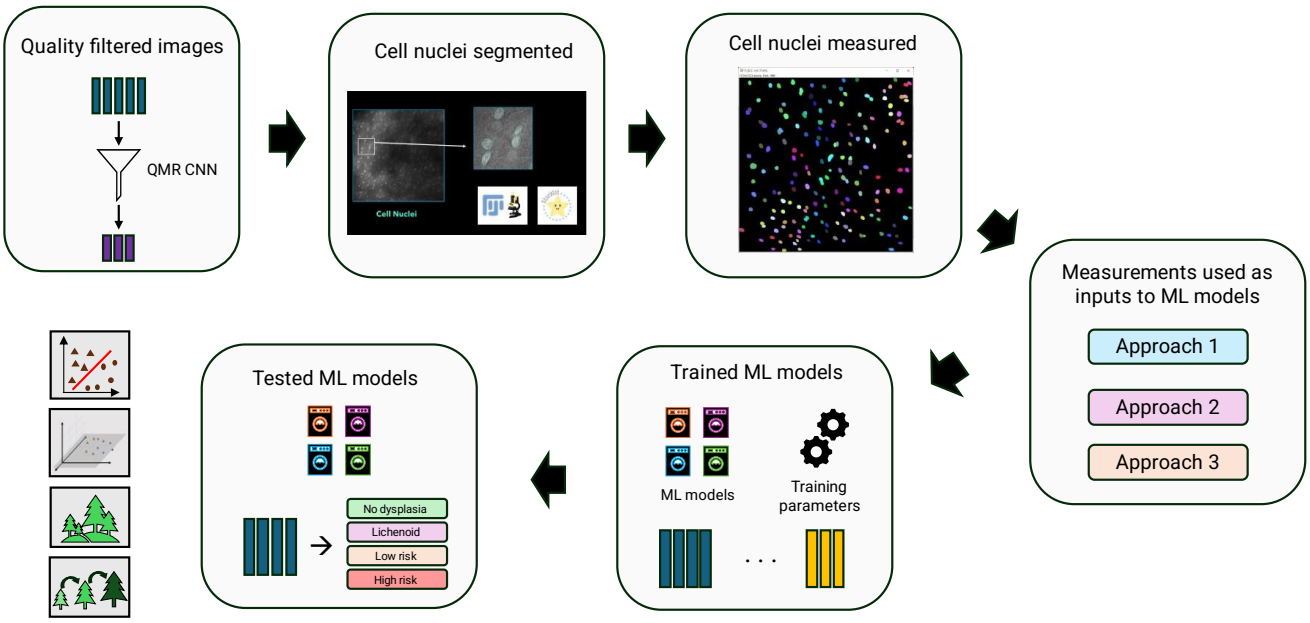


Figure 6.1. Overview of feature extraction methods used to obtain the results in this chapter

6.3. Segmentation model comparison results

The first step in preparation of the segmentation model involved a small experiment to compare to different segmentation tools. The two segmentation deep learning models based on the U-Net architecture Cellpose 2D model and the StarDist 2D model had very different results (Figure 6.2.).

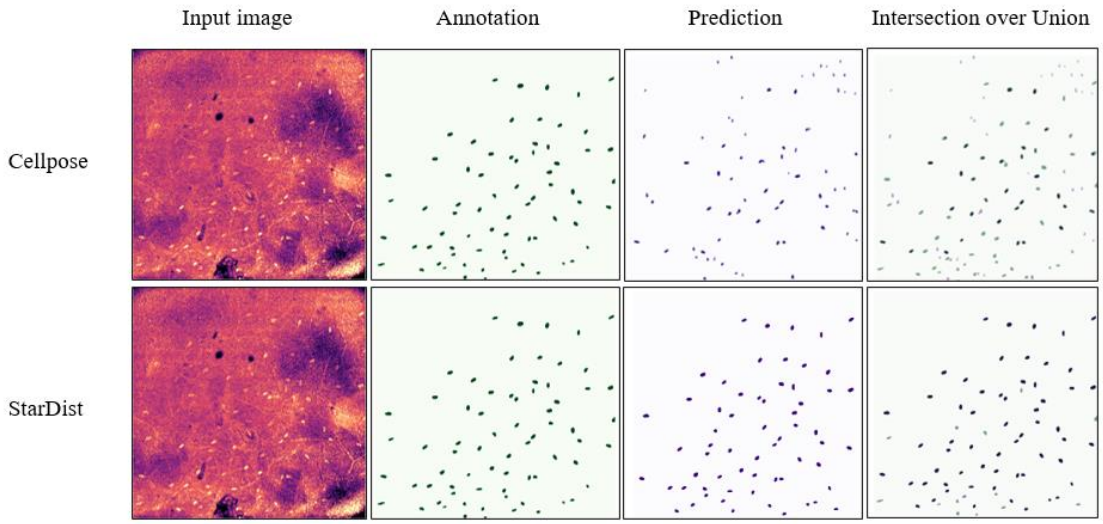


Figure 6.2. An example input OLP image with ground truth annotations, segmentation predictions, and intersection over union overlaps for both models

The Cellpose model performed poorly with an IoU of 0.16 and F1 score of 0.19. The StarDist model had a moderate performance at detecting nuclei with an IoU of 0.41 and F1 score of 0.56 (Table 6.1.). A power value of 0.232 was

Table 6.1. Results of the paired sample t-test to analyse the difference in performance between both models

	Cellpose model - mean (s.d.)	**StarDist model - mean (s.d.)**	**p = (t-test)**
IoU	0.163 (0.104)	0.413 (0.328)	**0.020**
Precision	0.188 (0.170)	0.596 (0.451)	**0.012**
Recall	0.275 (0.138)	0.523 (0.393)	0.101
F1 score	0.193 (0.139)	0.557 (0.420)	**0.015**

s.d. = standard deviation

estimated for the test dataset of 8 samples to detect an effect size of 0.5 with an alpha of 0.05.

The StarDist 2D model had better segmentation performance compared to the Cellpose 2D model in all metrics except recall with $p<0.05$ for IoU, precision and F1 score (Table 6.1.).

6.4. Experiment 2: Segmentation model development

After discarding images that were found to have no identifiable nuclei using a python script on the StarDist model the randomly selected dataset of 400 images, there were 362 images remaining with 16,219 annotated nuclei to be included in the StarDist 2D training and testing experiment. This consisted of 176 acriflavine images and 186 fluorescein images.

6.4.1. Training optimised models

The model was trained for 5,10 and 15 epochs to optimise the number of epochs. Epochs represent the number of times the model assesses every training sample. At the end of 5 epochs the training loss was 0.45 and the validation loss was 0.42Ad (Table 6.2.).

Table 6.2. Training and validation loss for StarDist 2D model trained for 5 epochs

Epochs	Training loss	Validation loss
1	1.10	1.01
2	1.01	0.98
3	0.97	0.79
4	0.61	0.53
5	0.45	0.42

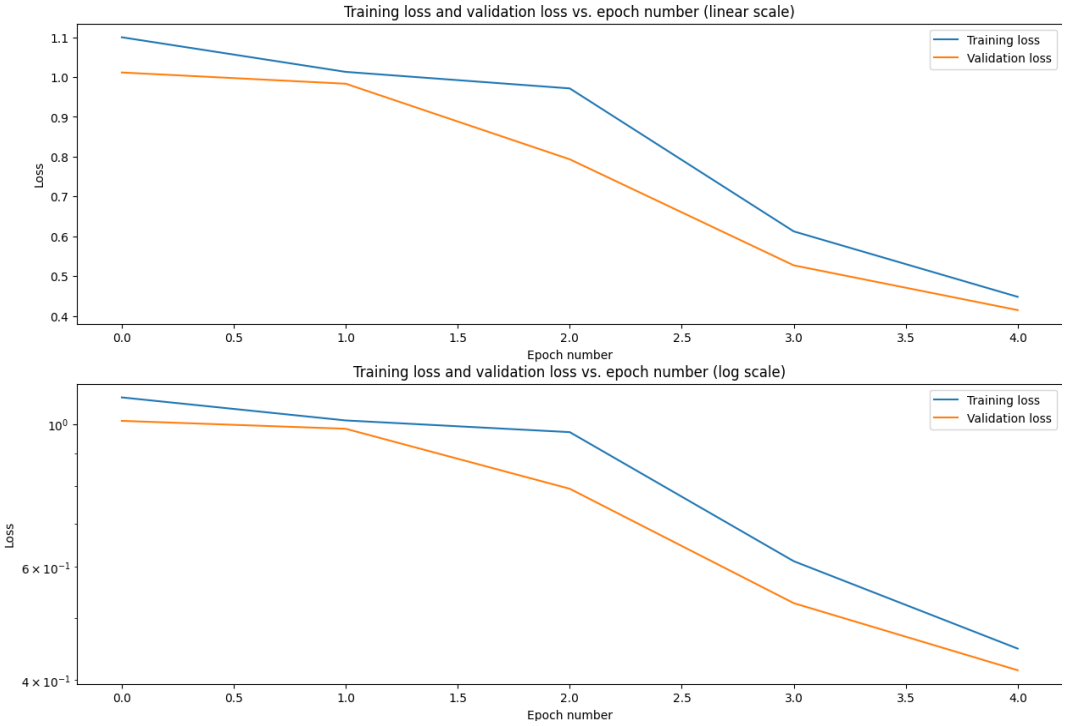


Figure 6.3. Training and validation loss graphs over the training cycle of 5 epochs (Top: linear scale, Bottom: logarithmic scale)

The training and validation curves showed both the training and validation loss decreasing over 5 epochs (Figure 6.3.).

Table 6.3. Training and validation loss for StarDist 2D model trained for 10 epochs

Epochs	Training loss	Validation loss
1	1.11	1.03
2	1.01	1.03
3	1.00	1.02
4	0.78	0.57
5	0.50	0.45
6	0.43	0.40
7	0.38	0.42
8	0.36	0.36
9	0.35	0.33
10	0.34	0.32

At the end of 10 epochs the training loss was 0.34 and the validation loss was 0.32 (Table 6.3.). The training and validation curves showed both the training and validation loss decreasing over 10 epochs and starting to stabilize close to the end of training (Figure 6.4.).

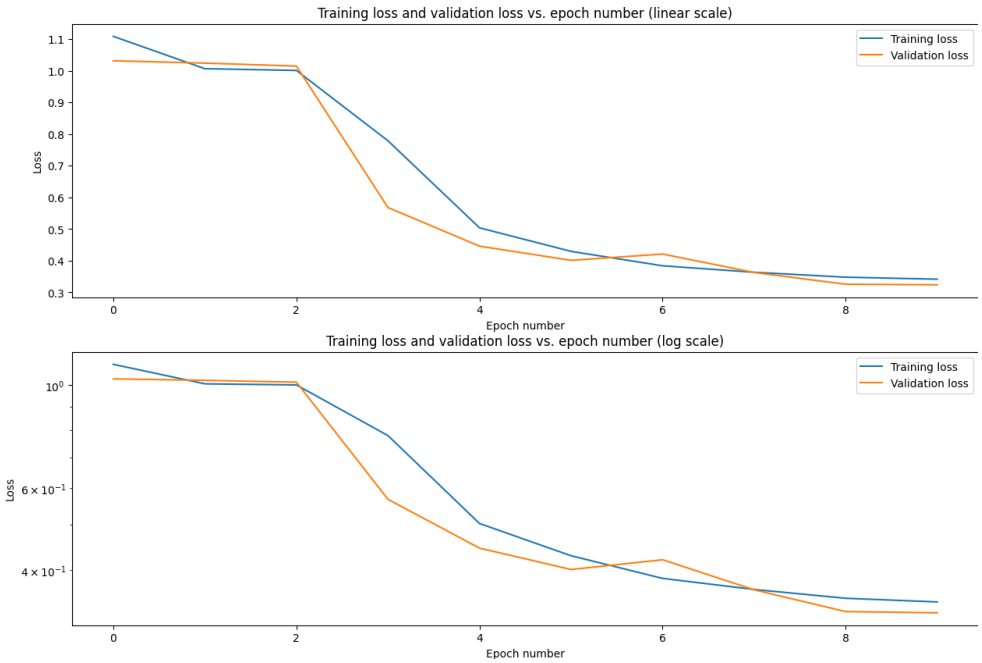


Figure 6.4. Training and validation loss graphs over the training cycle of 10 epochs (Top: linear scale, Bottom: logarithmic scale)

At the end of 15 epochs the training loss and validation loss were both 0.30 (Table 6.4.). As the training and validation loss stabilised (Figure 6.5.) at the end of 15 epochs the models were not trained for longer epoch cycles.

Table 6.4. Training and validation loss for StarDist 2D model trained for 15 epochs

Epochs	Training loss	Validation loss
1	1.10	1.05
2	1.01	1.05
3	1.00	1.03
4	0.73	0.55
5	0.47	0.42
6	0.40	0.39
7	0.37	0.39
8	0.35	0.34
9	0.34	0.31
10	0.33	0.32
11	0.32	0.31
12	0.32	0.36
13	0.31	0.30
14	0.30	0.326
15	0.30	0.30

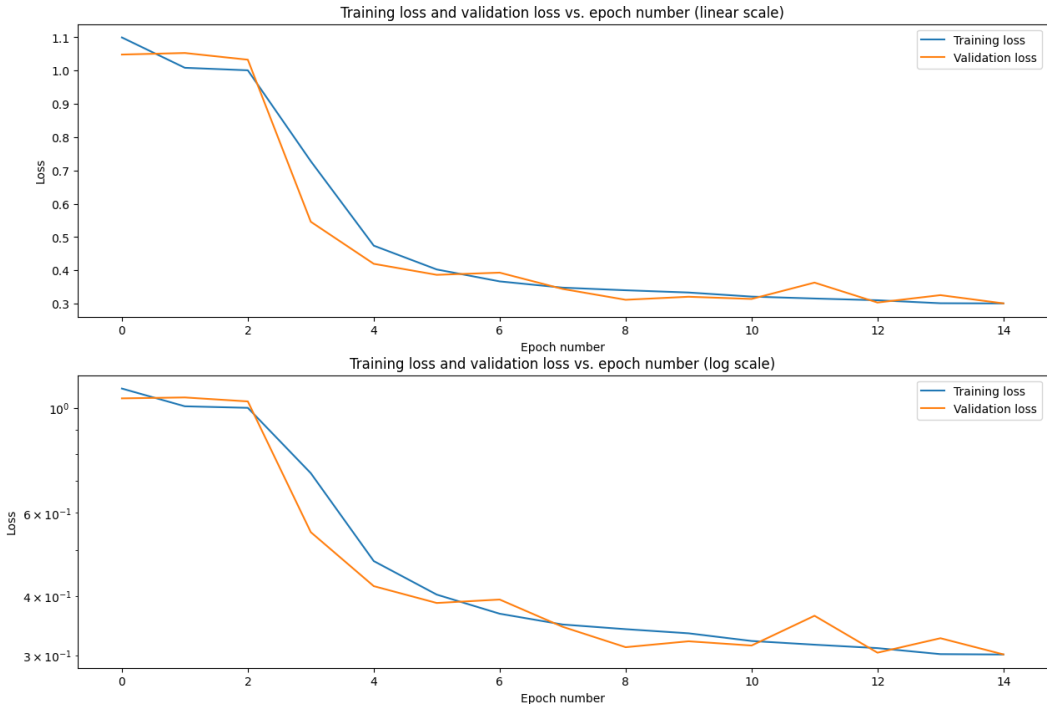


Figure 6.5. Training and validation loss graphs over the training cycle of 15 epochs (Top: linear scale, Bottom: logarithmic scale)

6.4.2. Segmentation performance

The 5, 10, and 15-epoch StarDist 2D models were evaluated with metrics such as accuracy, F1 score, mean true score (MTS), mean matched score (MMS) and panoptic quality (PQ) (Figure 6.6.).

The 10-epoch model had higher mean and median scores than both the 5- and 15-epoch models in accuracy, F1 score, MTS and PQ (Table 6.5.). However, the 15-epoch model had the highest mean and median MMS. This indicated that the 15-epoch model had better segmentation performance, but poorer detection accuracy compared to the 10-epoch model.

The 10-epoch model had the highest median accuracy (0.41), F1 score (0.58), MTD (0.49), and panoptic quality (0.472) while having the second highest MMS (0.82). The paired sample one-way ANOVA results (Table 6.5.) show that none of the metrics across all 3 models showed any statistically measurable difference except MMS.

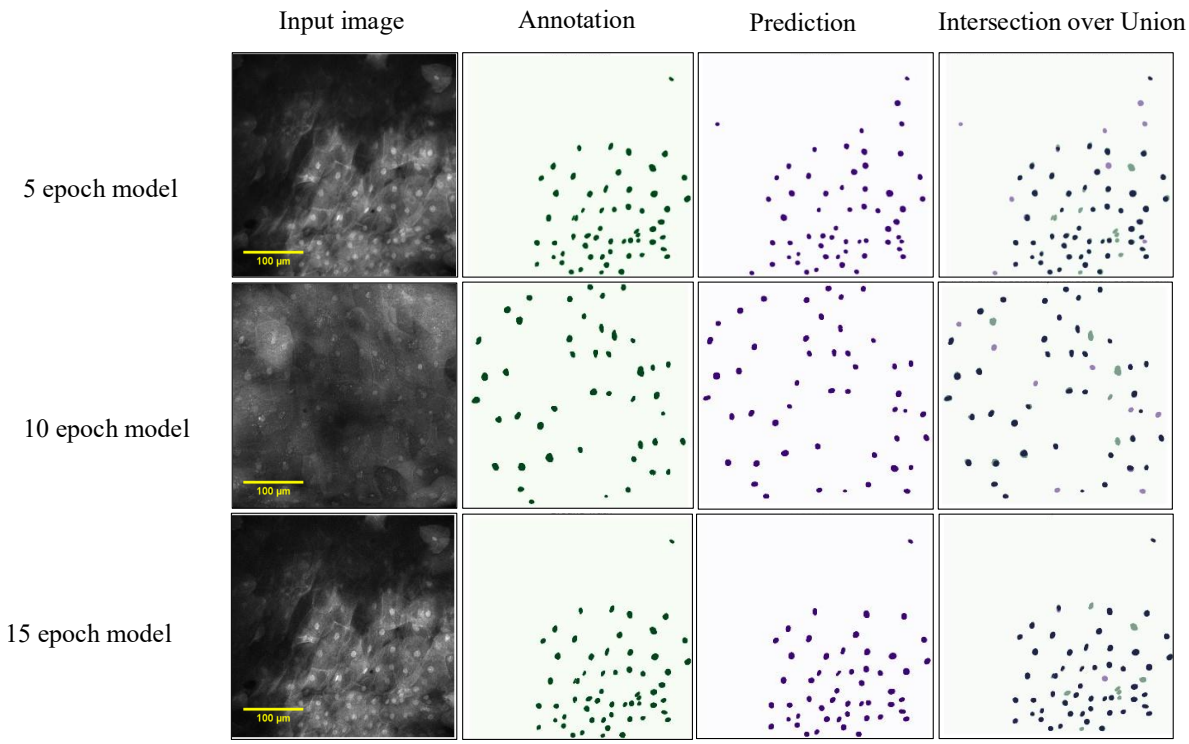


Figure 6.6. Example of Intersection over union (IoU) of the predictions over ground truth targets for the StarDist 2D models trained for 5, 10, and 15 epochs

Table 6.5. The summary statistics of the nuclei detection and segmentation of the 3 trained StarDist 2D models

Metric	Summary statistic	5-epoch model	10-epoch model	15-epoch model	P value for ANOVA
Accuracy	**Mean**	0.37	0.44	0.38	0.478
	Median	0.30	0.41	0.29	
	S.d.	0.24	0.25	0.24	
F1 score	**Mean**	0.49	0.56	0.51	0.505
	Median	0.47	0.58	0.46	
	S.d.	0.26	0.26	0.25	
Mean true score	**Mean**	0.39	0.49	0.39	0.108
	Median	0.39	0.49	0.31	
	S.d.	0.19	0.23	0.22	
Mean matched score	**Mean**	0.72	0.79	0.81	**0.049**
	Median	0.75	0.82	0.85	
	S.d.	0.14	0.15	0.16	
Panoptic quality	**Mean**	0.37	0.46	0.43	0.196
	Median	0.35	0.47	0.38	
	S.d.	0.20	0.22	0.22	

6.4.3. Nuclei measurements

The entire dataset of diagnostic quality micrographs consisted of 1343 acriflavine and 640 fluorescein images as filtered by the QMR CNN described in Chapter 4.

6.4.3.1. Measurements on the acriflavine dataset

The diagnostic category distribution of acriflavine dataset images (n=1343) is depicted in Table 6.6. The StarDist 2D model predicted 91,550 nuclei across

1322 images. A total of 21 images were excluded for having zero nuclei identified (Table 6.6.).

Table 6.6. Distribution of acriflavine nuclei segmented by the trained StarDist 2D across all diagnostic categories

Disease category	Images included	Nuclei identified
High-risk	86	**3072**
high-grade dysplasia	73	2982
OSCC	13	90
Lichenoid	340	**24736**
lichenoid inflammation	215	16617
oral lichen planus	125	8119
Low-risk	350	**31786**
atypia	50	8045
low-grade dysplasia	162	21034
verrucous hyperplasia	138	2707
No dysplasia	546	**31956**
amalgam tattoo	31	3195
chronic inflammation	19	1580
denture associated gingival hyperplasia	29	851
histopathologically normal tissue	2	20
fibroepithelial polyp	2	4293
focal papillomatosis	47	4
hyperplasia & hyperkeratosis	398	21762
squamous papilloma	6	14
verruciform xanthoma	12	237
Grand Total	**1322**	**91550**

Nuclei from images with no dysplasia had the lowest median integrated density indicating they had relatively low fluorescence intensity overall (Table 6.7.). Lichenoid nuclei had the highest median pixel area while also having lowest mean pixel intensity and intensity standard deviation (Table 6.7.).

Table 6.7. Summary statistics of the measurements of nuclei segmented in acriflavine images

		no dysplasia	lichenoid	low-risk	high-risk	p value (ANOVA)
area	**mean**	386.08	397.98	376.88	380.90	**0.046**
	median	371.58	390.62	367.71	361.00	
	s.d.	105.54	90.77	83.46	72.42	
mean intensity	**mean**	120.87	119.34	135.07	143.42	**<0.001**
	median	117.13	113.40	130.89	143.93	
	s.d.	40.86	38.37	34.62	34.89	
intensity standard deviation	**mean**	20.81	19.18	24.06	20.37	**<0.001**
	median	19.26	18.00	23.37	19.58	
	s.d.	6.72	5.54	6.54	4.19	
circularity	**mean**	0.93	0.94	0.93	0.94	**<0.001**
	median	0.93	0.94	0.94	0.94	
	s.d.	0.02	0.01	0.01	0.01	
aspect ratio	**mean**	1.36	1.33	1.34	1.33	**<0.001**
	median	1.35	1.31	1.34	1.33	
	s.d.	0.07	0.08	0.08	0.06	
integrated density	**mean**	46695.92	47501.86	50724.53	54893.17	**0.003**
	median	42679.33	44130.67	48478.73	50209.73	
	s.d.	21470.96	19766.56	17874.42	18702.95	
s.d. = standard deviation, Units for area = pixel^2 All other metrics in this table are unitless.						

Low-risk nuclei had the highest median intensity standard deviation indicating the large variability in nuclei brightness (Table 6.7.). High-risk nuclei had the highest median mean intensity and integrated density indicating these nuclei were the brightest (Table 6.7.).

The circularity and aspect ratio values were relatively even across all diagnostic categories indicating all segmented nuclei were relatively even circles (Table 6.7.). The one-way ANOVA showed a significant difference between all measurement metrics across the diagnostic categories (Table 6.7.).

The Tukey's pairwise comparison post-hoc test showed significant differences between specific diagnostic category pairs (Table 6.8.). The nuclei area measurement only showed significant differences between the lichenoid and low-risk nuclei ($p=0.029$).

Mean pixel intensity values for high-risk and low-risk nuclei were individually significantly different from no dysplasia and lichenoid nuclei ($p<0.005$). Standard deviation of pixel intensity for normal, lichenoid and low-risk nuclei was significantly different compared to each other and high-risk values were significantly different from low-risk ($p<0.005$) (Table 6.8.).

Table 6.8: Tukey's pairwise comparison post hoc test for ANOVA for the acriflavine nuclei measurements

	Category 1	**Category 2**	**Mean difference**	**Lower 95% confidence interval**	**Upper 95% confidence interval**	**p value**
area	high-risk	lichenoid	17.08	-17.78	51.94	0.588
	high-risk	low-risk	-4.02	-39.01	30.97	0.991
	high-risk	normal	5.18	-29.00	39.35	0.980
	lichenoid	low-risk	-21.10	-40.66	-1.54	**0.029**
	lichenoid	normal	-11.90	-29.96	6.16	0.327
	low-risk	normal	9.20	-9.12	27.52	0.568
mean brightness	high-risk	lichenoid	-24.08	-38.24	-9.93	**<0.001**
	high-risk	low-risk	-8.35	-22.56	5.86	0.430
	high-risk	normal	-22.56	-36.43	-8.68	**<0.001**
	lichenoid	low-risk	15.73	7.79	23.67	**<0.001**

	lichenoid	normal	1.53	-5.81	8.86	0.950
	low-risk	normal	-14.20	-21.64	-6.76	**<0.001**
s.d. of brightness	high-risk	lichenoid	-1.19	-3.50	1.12	0.549
	high-risk	low-risk	3.69	1.37	6.01	**<0.001**
	high-risk	normal	0.44	-1.83	2.71	0.959
	lichenoid	low-risk	4.88	3.58	6.17	**<0.001**
	lichenoid	normal	1.63	0.43	2.83	**0.003**
	low-risk	normal	-3.25	-4.46	-2.03	**<0.001**
circularity	high-risk	lichenoid	0.00	-0.01	0.01	1.000
	high-risk	low-risk	0.00	-0.01	0.00	0.380
	high-risk	normal	-0.01	-0.01	0.00	**0.029**
	lichenoid	low-risk	0.00	-0.01	0.00	**0.016**
	lichenoid	normal	-0.01	-0.01	0.00	**<0.001**
	low-risk	normal	0.00	-0.01	0.00	0.152
aspect ratio	high-risk	lichenoid	-0.01	-0.03	0.02	0.954
	high-risk	low-risk	0.01	-0.02	0.04	0.815
	high-risk	normal	0.03	0.00	0.05	**0.057**
	lichenoid	low-risk	0.02	0.00	0.03	**0.060**
	lichenoid	normal	0.03	0.02	0.05	**0.000**
	low-risk	normal	0.02	0.00	0.03	**0.013**
integrated density	high-risk	lichenoid	-7391.31	-14764.81	-17.81	**0.049**
	high-risk	low-risk	-4168.64	-11570.50	3233.22	0.469
	high-risk	normal	-8197.24	-15426.37	-968.12	**0.019**
	lichenoid	low-risk	3222.68	-915.59	7360.94	0.187
	lichenoid	normal	-805.93	-4626.67	3014.81	0.949
	low-risk	normal	-4028.61	-7903.80	-153.42	**0.038**

s.d. = standard deviation

Nucleus circularity values for no dysplasia and lichenoid were significantly different from each other while no dysplasia was significantly different from high-risk and lichenoid was significantly different from low-risk. No dysplasia nuclei had significantly different aspect ratio values compared to lichenoid and low-risk nuclei. Integrated density of no dysplasia and high-risk nuclei was significantly different from each other while high-risk vs lichenoid values and no dysplasia vs low-risk values were also significantly different (Table 6.8.). Nuclear characteristics varied significantly across diagnostic categories in this analysis. No dysplasia nuclei exhibited the lowest overall fluorescence intensity, while high-risk nuclei were the brightest with the highest mean intensity and integrated density.

Lichenoid nuclei were notable for their large size but low intensity, and low-risk nuclei showed the greatest variability in brightness. Statistical analysis, including ANOVA and Tukey's post-hoc tests, confirmed significant differences in multiple nuclear metrics between diagnostic groups, suggesting potential for these measurements in distinguishing between different stages of oral epithelial dysplasia (Table 6.8.).

6.4.3.2. Measurements on the fluorescein dataset

The StarDist 2D model predicted 10,722 nuclei across 640 images (Table 6.9.). A total of 337 images were excluded for having less than 10 nuclei.

Table 6.9. Distribution of fluorescein nuclei segmented by the trained StarDist 2D across all diagnostic categories

Disease category	Images included	Nuclei identified
high-risk	**42**	**1031**
high-grade dysplasia	42	1031
OSCC	0	0
lichenoid	**107**	**4251**
lichenoid inflammation	66	2911
oral lichen planus	41	1340
low-risk	**66**	**1488**
low-grade dysplasia	58	1370
verrucous hyperplasia	8	118
normal	**88**	**2350**
chronic inflammation	5	120
denture associated gingival hyperplasia	2	43
fibroepithelial polyp	4	69
focal papillomatosis	8	229
hyperplasia & hyperkeratosis	69	1889
squamous papilloma	0	0
Grand Total	**303**	**9120**

No dysplasia nuclei had the highest median mean intensity while lichenoid nuclei had the lowest values (Table 6.10.). Lichenoid nuclei also had the lowest intensity standard deviation and integrated density indicating that these nuclei had the lowest fluorescence brightness (Table 6.10.). Low-risk nuclei had the lowest median area indicating these nuclei were smaller in general (Table 6.10.). Notably high-risk nuclei had the highest median area, intensity standard deviation and integrated density indicating these nuclei were on average larger and brighter (Table 6.10.). The circularity and aspect ratio

values were relatively even across all diagnostic categories indicating all segmented nuclei were relatively even circles (Table 6.10.).

Table 6.10. Summary statistics of the measurements of nuclei segmented in fluorescein images

		normal	lichenoid	low-risk	high-risk	p value (ANOVA)
area	**mean**	361.28	377.27	339.14	443.98	**<0.001**
	median	341.44	353.38	312.16	422.52	
	s.d.	73.70	103.74	100.51	82.30	
mean brightness	**mean**	155.70	133.51	153.59	148.83	**<0.001**
	median	154.12	126.12	153.54	150.14	
	s.d.	29.70	36.28	28.28	25.70	
s.d. of brightness	**mean**	18.20	18.23	20.55	23.52	**<0.001**
	median	17.43	16.29	20.46	24.18	
	s.d.	5.00	6.04	4.45	4.30	
circularity	**mean**	0.934	0.939	0.940	0.937	**0.03**
	median	0.939	0.942	0.941	0.936	
	s.d.	0.017	0.011	0.014	0.011	
aspect ratio	**mean**	1.352	1.320	1.310	1.338	**<0.001**
	median	1.340	1.314	1.298	1.343	
	s.d.	0.087	0.059	0.074	0.063	
integrated density	**mean**	56701.55	50463.23	52962.00	67987.08	**<0.001**
	median	55853.89	46706.04	49147.80	64526.38	
	s.d.	17284.41	20465.11	20844.34	20670.57	

Units: Area – $pixel^2$

All other metrics in this table are unitless.

The one-way ANOVA for fluorescein nuclei measurements was statistically significant for all measurement metrics except for circularity (Table 6.10.). The Tukey's pairwise comparison post-hoc test showed a variable combination of significant differences between pairs of diagnostic categories (Table 6.11.).

Nucleus area values for high-risk nuclei were significantly different compared to all other diagnostic categories while lichenoid and low-risk values were also significantly different. For mean nucleus pixel intensity measurements, the Lichenoid nuclei values were significantly different compared to all other diagnostic categories. For standard deviation of pixel intensity all combinations of pairs were significantly different to each other except for no dysplasia vs lichenoid. Aspect ratio values of no dysplasia nuclei were significantly different to lichenoid and low-risk nuclei. Integrated density values of high-risk nuclei were significantly different from all other diagnostic categories (Table 6.11.).

No dysplasia nuclei exhibited the highest median mean intensity, while lichenoid nuclei had the lowest fluorescence brightness, as indicated by their low intensity standard deviation and integrated density. High-risk nuclei stood out with the largest median area, highest intensity standard deviation, and greatest integrated density, suggesting they were generally larger and brighter. Statistical analysis, including ANOVA and Tukey's post-hoc tests, revealed significant differences in multiple nuclear metrics between diagnostic groups, with high-risk nuclei often distinguishing themselves from other categories, particularly in terms of area and integrated density (Table 6.11.).

Table 6.11. Tukey's pairwise comparison post hoc test for ANOVA for the fluorescein nuclei measurements

	Category 1	Category 2	Mean difference	Lower 95% confidence interval	Upper 95% confidence interval	p value
area	high-risk	lichenoid	-66.71	-110.13	-23.28	**0.001**
	high-risk	low-risk	-104.84	-151.91	-57.77	**<0.001**
	high-risk	normal	-82.70	-127.43	-37.97	**<0.001**
	lichenoid	low-risk	-38.13	-75.46	-0.80	**0.043**
	lichenoid	normal	-15.99	-50.31	18.33	0.625
	low-risk	normal	22.14	-16.69	60.97	0.455
mean brightness	high-risk	lichenoid	-15.32	-30.10	-0.53	**0.039**
	high-risk	low-risk	4.76	-11.27	20.79	0.869
	high-risk	normal	6.87	-8.36	22.10	0.649
	lichenoid	low-risk	20.07	7.36	32.78	**<0.001**
	lichenoid	normal	22.19	10.50	33.87	**<0.001**
	low-risk	normal	2.11	-11.11	15.33	0.976
s.d. of brightness	high-risk	lichenoid	-5.29	-7.73	-2.84	**<0.001**
	high-risk	low-risk	-2.97	-5.62	-0.32	**0.021**
	high-risk	normal	-5.31	-7.83	-2.79	**<0.001**
	lichenoid	low-risk	2.32	0.21	4.42	**0.024**
	lichenoid	normal	-0.03	-1.96	1.91	1.000
	low-risk	normal	-2.34	-4.53	-0.16	**0.031**
circularity	high-risk	lichenoid	0.002	-0.005	0.008	0.873
	high-risk	low-risk	0.003	-0.005	0.010	0.797
	high-risk	normal	-0.003	-0.010	0.004	0.614
	lichenoid	low-risk	0.001	-0.005	0.006	0.993
	lichenoid	normal	-0.005	-0.010	0.000	0.054

	low-risk	normal	-0.006	-0.011	0.000	0.060
aspect ratio	high-risk	lichenoid	-0.018	-0.052	0.015	0.497
	high-risk	low-risk	-0.029	-0.065	0.008	0.185
	high-risk	normal	0.014	-0.021	0.049	0.730
	lichenoid	low-risk	-0.010	-0.039	0.019	0.802
	lichenoid	normal	0.032	0.006	0.059	**0.011**
	low-risk	normal	0.043	0.012	0.073	**0.002**
integrated density	high-risk	lichenoid	-17523.85	-26795.52	-8252.18	**<0.001**
	high-risk	low-risk	-15025.08	-25075.79	-4974.37	**0.001**
	high-risk	normal	-11285.53	-20835.17	-1735.89	**0.013**
	lichenoid	low-risk	2498.77	-5470.90	10468.45	0.850
	lichenoid	normal	6238.32	-1089.34	13565.99	0.126
	low-risk	normal	3739.55	-4551.86	12030.96	0.649

s.d. = standard deviation

6.5. Experiment 3: Machine learning of extracted nuclear features for diagnostic triage

6.5.1. Acriflavine ML performance results

All measurement data across the 6 measurements of mean pixel intensity, standard deviation of pixel intensity, mean surface area, integrated density, circularity and aspect ratio of all 91,550 acriflavine stained nuclei was standardised using the z-score method to have a mean of 0 and a standard deviation of 1. There were 4 ML models developed for each of the 4 model types across all approaches.

6.5.1.1. **Approach 1: Direct classification of measurement means for size, shape and brightness.**

The SVM model emerged as the best performer in classifying images based on the mean nuclei measurements per image, achieving the highest overall performance across the different diagnostic categories (Table 6.12. & 6.13.).

For the no dysplasia category, the SVM model achieved an accuracy of 0.64, sensitivity of 0.60, and F1-score of 0.55, indicating reasonably good classification of no dysplasia nuclei. The lichenoid category also showed promising results, with the SVM model achieving an accuracy of 0.70, sensitivity of 0.43, and F1-score of 0.46 (Table 6.12. & 6.13.).

Table 6.12. Approach 1 test results for all 4 ML models with respect to each diagnostic category (1 vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	0.58	0.64	0.64	0.58
	sensitivity	0.68	0.59	0.57	0.76
	specificity	0.52	0.67	0.68	0.47
	precision	0.46	0.52	0.52	0.46
	f1score	0.55	0.55	0.54	0.58
lichenoid	accuracy	0.72	0.70	0.70	0.74
	sensitivity	0.42	0.43	0.55	0.37
	specificity	0.85	0.81	0.76	0.89
	precision	0.54	0.49	0.49	0.60
	f1score	0.47	0.46	0.52	0.45
low-risk	accuracy	0.70	0.75	0.72	0.69
	sensitivity	0.36	0.55	0.45	0.30
	specificity	0.84	0.83	0.83	0.84
	precision	0.46	0.55	0.50	0.43
	f1score	0.40	0.55	0.47	0.35
high-risk	accuracy	0.95	0.94	0.94	0.95
	sensitivity	0	0.20	0	0
	specificity	1	0.98	0.99	1
	precision	0	0.33	0	0
	f1score	0	0.25	0	0

The low-risk category had the highest performance, with the SVM model reaching an accuracy of 0.75, sensitivity of 0.55, and F1-score of 0.55. However, the high-risk category proved to be the most challenging, with the SVM model struggling to achieve high sensitivity (0.20) and precision (0.33), resulting in a low F1-score of 0.25 (Table 6.12. & 6.13.).

Table 6.13. Approach 1 ranks for all 4 ML models with respect to each diagnostic category (1vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	3	1	1	4
	sensitivity	2	3	4	1
	specificity	3	2	1	4
	precision	3	2	1	3
	f1score	2	3	4	1
lichenoid	accuracy	2	3	3	1
	sensitivity	3	2	1	4
	specificity	2	3	4	1
	precision	2	4	3	1
	f1score	2	3	1	4
low-risk	accuracy	3	1	2	4
	sensitivity	3	1	2	4
	specificity	2	3	3	1
	precision	3	1	2	4
	f1score	3	1	2	4
high-risk	accuracy	1	3	3	1
	sensitivity	2	1	2	2
	specificity	1	4	3	1
	precision	2	1	2	2
	f1score	2	1	2	2
Aggregate rank (sum)		46	43	46	49
Final rank		2	1	2	4

6.5.1.2. Approach 2: Clustering and feature selection for size, shape and brightness

The elbow plot of within cluster sum of squares (WCSS) that is the sum of the squared distance between each point and the centroid of its cluster versus number of clusters (k) showed a sharp decline in WCSS for initial values of k, followed by a gentler slope as 'k' was progressively increased (Figure 6.7.). Thus, the clusters of nuclei measurement data become more cohesive as the number of clusters (k) is increased, however that also increases the complexity of the model, which can significantly increase computational costs and time.

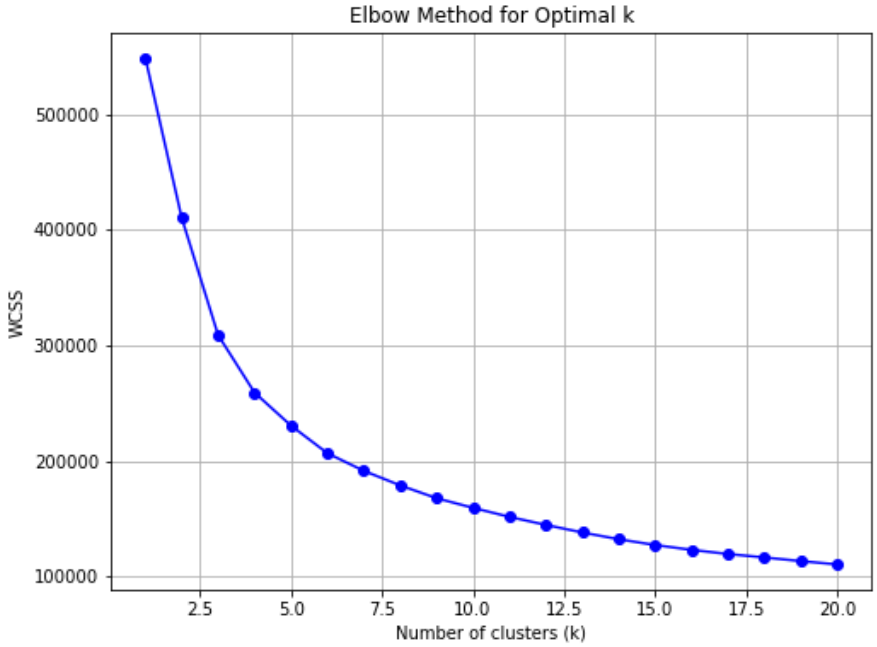


Figure 6.7. Elbow plot of WCSS vs number of clusters for acriflavine nuclei measurements

The curve initially drops steeply as the number of clusters increase until k=4, k=5, k=6 and k=7. This sharp decrease suggests adding more clusters improved the compactness of cluster groups. After k=6, the rate of decease slows forming a gradual slope. Thus, adding more clusters beyond this point doesn't significantly improve grouping of objects. The plot suggests a good balance between number of clusters and compactness of clusters can be found around k=4, k=5, k=6 and k=7 (Figure 6.7.).

Non-linear relationships

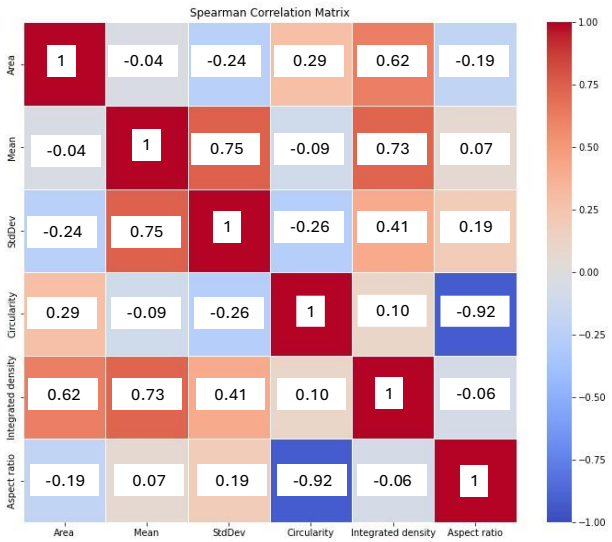


Figure 6.8. Spearman's correlation matrix for acriflavine nuclei measurements

A spearman's correlation matric shows the relationships between different measurement metrics and helps identify the metrics that correlate strongly with each other and thus could potentially introduce noise reducing the distinctness of the diagnostic groups that the nuclei belong to (Figure 6.8.). The correlation matrix shows area and mean have a fairly strong positive correlation with integrated density. Mean shows a fairly strong positive correlation with standard deviation of pixel intensity and circularity shows a strong negative correlation with aspect ratio (Figure 6.8.).

Based on the correlation analysis we can avoid multicollinearity by eliminating one half of highly correlated feature pairs. These features do not correlate with each other and should therefore provide unique and diverse information when developing clustering and classification algorithms.

Three features were eliminated in a step-wise manner when comparing pairs of features with the highest correlation (Figure 6.9.). At each step a new feature set was created beginning with the original 6-feature set.

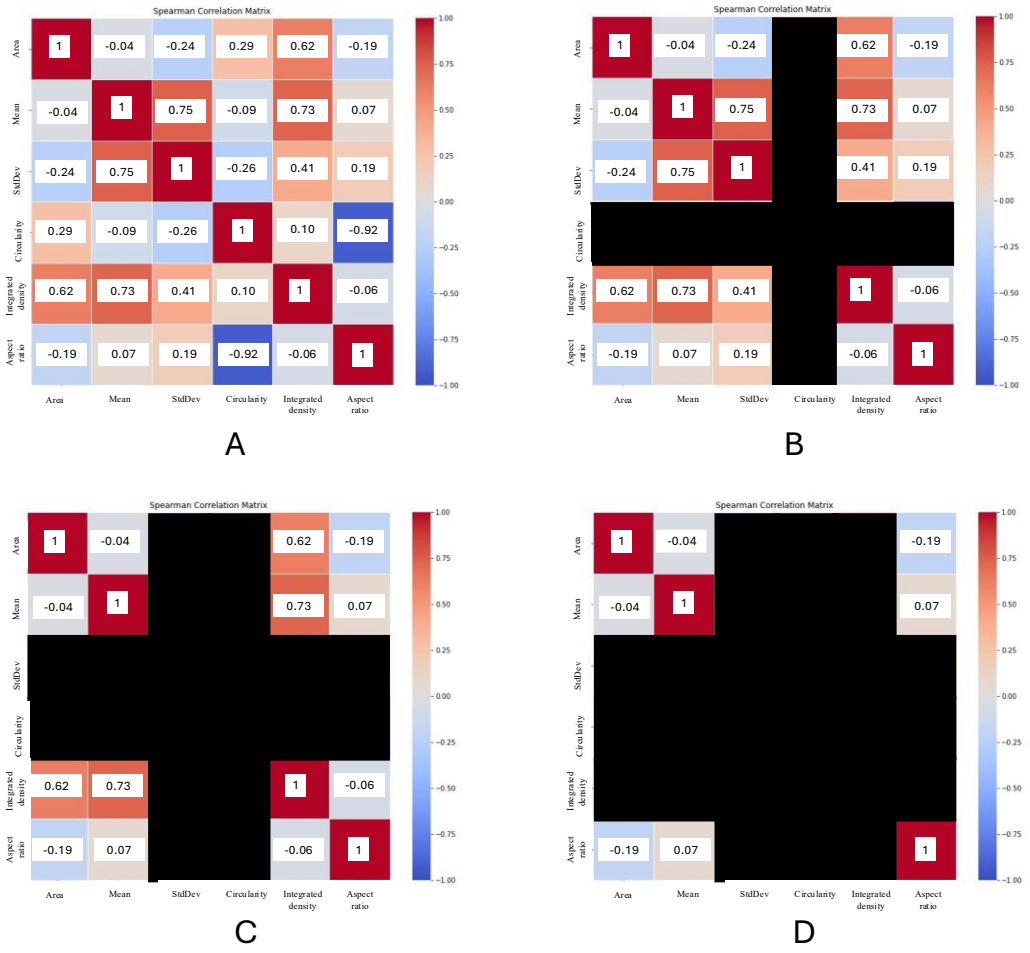


Figure 6.9. Spearman's correlation matrix of the 6 measurement features of the nuclei in acriflavine images. The four panels show which features were eliminated at each step of the feature selection. A) original 6-feature set, B) 5-feature set, C) 4-feature set, D) 3-feature set

The first pair to have the highest absolute correlation were circularity-aspect ratio (0.92). The sum of absolute correlation of both these metrics against the rest:

- Circularity = 0.10 + 0.26 + 0.09 + 0.29 = 0.76
- Aspect ratio = 0.06 + 0.19 + 0.07 + 0.19 = 0.51

The absolute sum of correlation for aspect ratio against all other metrics is lower (0.51), indicating it is less correlated with the other metrics overall compared to circularity. Thus, circularity was eliminated and aspect ratio retained. The 5-feature set comprised of the remaining features of mean pixel intensity, area, standard deviation of pixel intensity, integrated density, and aspect ratio (Figure 6.9.B).

Mean pixel intensity and standard deviation of pixel intensity had the next highest absolute correlation (0.746). The sum of absolute correlation of both these metrics against the rest:

- Mean pixel intensity = 0.068 + 0.732 + 0.036 = 0.836
- Standard deviation of pixel intensity = 0.194 + 0.411 + 0.244 = 0.849

The absolute sum of correlation for mean pixel intensity against all other metrics is slightly lower (0.836), indicating it is less correlated with the other metrics overall compared to standard deviation of pixel intensity. Thus, standard deviation of pixel intensity was eliminated and mean pixel intensity retained. The 4-feature set comprised of the remaining features of mean pixel intensity, area, integrated density, and aspect ratio (Figure 6.9.C).

Mean pixel intensity and integrated density have the next highest absolute correlation (0.73) The sum of absolute correlation of both these metrics against the rest:

- Mean pixel intensity = 0.07 + 0.04 = 0.11
- Integrated density = 0.62 + 0.06 = 0.68

The absolute sum of correlation for mean pixel intensity against all other metrics is lower (0.11), indicating it is less correlated with the other metrics overall compared to integrated density. Thus, integrated density was eliminated and mean pixel intensity retained. The 3-feature set comprised of the remaining features of mean pixel intensity, area, and aspect ratio (Figure 6.9.D).

ML test results for approach 2

Twenty logistic regression models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. The models across all combinations had very poor results in correctly identifying high-risk images. They also performed poorly with lichenoid and low-risk images while having a moderate level of success with no dysplasia images. Upon ranking, the best performing model was the one developed on the 5-feature set with k=4 for measurement clustering (Table 6.14.).

Table 6.14. Test results of all logistic regression models developed across all combinations of feature sets and clustering for all diagnostic categories on acriflavine images

Combination		Normal					Lichenoid					Low-risk					High-risk					Model rank
Features	Clusters (k)	Acc	Sen	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	
6	3	0.45	0.96	0.09	0.43	0.59	0.73	0.11	0.94	0.39	0.17	0.74	0.00	1.00	0.00	0.00	0.93	0	1	0	0	9
6	4	0.50	0.80	0.29	0.44	0.57	0.73	0.14	0.93	0.41	0.21	0.69	0.24	0.86	0.37	0.29	0.93	0	1	0	0	5
6	5	0.51	0.76	0.33	0.45	0.56	0.72	0.18	0.91	0.40	0.25	0.70	0.27	0.86	0.40	0.32	0.93	0	1	0	0	4
6	6	0.48	0.76	0.28	0.42	0.54	0.72	0.15	0.91	0.37	0.22	0.70	0.24	0.87	0.39	0.29	0.93	0	1	0	0	14
6	7	0.50	0.80	0.29	0.44	0.57	0.73	0.12	0.94	0.40	0.19	0.71	0.28	0.87	0.43	0.34	0.93	0	1	0	0	2
5	3	0.48	0.87	0.21	0.43	0.58	0.74	0.00	1.00	0.00	0.00	0.71	0.28	0.86	0.42	0.34	0.93	0	1	0	0	6
5	4	0.55	0.81	0.36	0.47	0.60	0.74	0.18	0.93	0.46	0.26	0.70	0.29	0.85	0.41	0.34	0.93	0	1	0	0	1
5	5	0.46	0.77	0.24	0.42	0.54	0.71	0.11	0.92	0.32	0.16	0.71	0.24	0.88	0.42	0.30	0.93	0	1	0	0	15
5	6	0.48	0.81	0.25	0.43	0.56	0.73	0.08	0.96	0.39	0.13	0.71	0.28	0.87	0.43	0.34	0.93	0	1	0	0	3
5	7	0.50	0.76	0.33	0.44	0.56	0.70	0.06	0.92	0.20	0.09	0.72	0.37	0.84	0.46	0.41	0.93	0	1	0	0	10
4	3	0.42	0.98	0.03	0.42	0.58	0.74	0.03	0.98	0.40	0.06	0.74	0.02	1.00	0.50	0.03	0.93	0	1	0	0	8
4	4	0.45	0.83	0.19	0.42	0.56	0.75	0.09	0.97	0.55	0.16	0.69	0.16	0.87	0.31	0.21	0.93	0	1	0	0	12

4	5	0.41	0.79	0.15	0.39	0.53	0.74	0.08	0.96	0.42	0.13	0.69	0.15	0.88	0.31	0.20	0.93	0	1	0	0	17
4	6	0.41	0.81	0.13	0.40	0.53	0.72	0.03	0.96	0.22	0.05	0.70	0.16	0.89	0.36	0.22	0.93	0	1	0	0	19
4	7	0.42	0.73	0.21	0.39	0.51	0.70	0.14	0.89	0.30	0.19	0.69	0.13	0.89	0.30	0.18	0.93	0	1	0	0	20
3	3	0.41	0.92	0.05	0.40	0.56	0.73	0.06	0.96	0.33	0.10	0.72	0.02	0.98	0.20	0.03	0.93	0	1	0	0	17
3	4	0.44	0.94	0.09	0.42	0.58	0.73	0.00	0.98	0.00	0.00	0.73	0.10	0.95	0.44	0.17	0.93	0	1	0	0	13
3	5	0.47	0.84	0.21	0.43	0.57	0.72	0.17	0.92	0.41	0.24	0.72	0.12	0.93	0.38	0.18	0.93	0	1	0	0	7
3	6	0.46	0.80	0.22	0.42	0.55	0.72	0.17	0.91	0.39	0.23	0.71	0.15	0.92	0.39	0.21	0.93	0	1	0	0	10
3	7	0.44	0.82	0.17	0.41	0.55	0.73	0.09	0.95	0.38	0.15	0.69	0.13	0.89	0.31	0.19	0.93	0	1	0	0	16
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																					

Twenty SVM models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. Similar to the logistic regression models the SVM models also performed exceptionally poorly with the high-risk images with not identifying a single image correctly in a majority of the cases. The models also performed poorly with the lichenoid and low-risk images while having a moderate performance in detecting no dysplasia oral mucosa micrographs. The best performing SVM model using this analysis approach was the original 6-feature set with k=7 for measurement clustering (Table 6.15.).

Table 6.15. Test results of all SVM models developed across all combinations of feature sets and clustering for all diagnostic categories on acriflavine images

Combination		**Normal**					**Lichenoid**					**Low-risk**					**High-risk**					**Model rank**
Features	**Clusters (k)**	**Acc**	**Sen**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	
6	3	0.49	0.88	0.23	0.44	0.59	0.74	0.00	1.00	0.00	0.00	0.66	0.21	0.83	0.30	0.24	0.93	0	1	0	0	18
6	4	0.57	0.78	0.42	0.49	0.60	0.73	0.09	0.95	0.40	0.15	0.70	0.46	0.78	0.43	0.44	0.93	0	1	0	0	15

6	5	0.61	0.76	0.51	0.52	0.62	0.71	0.24	0.87	0.40	0.30	0.72	0.44	0.83	0.48	0.46	0.93	0	1	0	0	8
6	6	0.62	0.71	0.56	0.53	0.61	0.71	0.38	0.83	0.43	0.40	0.75	0.44	0.86	0.53	0.48	0.93	0	1	0	0	2
6	7	0.67	0.79	0.58	0.57	0.66	0.72	0.32	0.86	0.45	0.37	0.75	0.49	0.84	0.52	0.50	0.93	0	1	0	0	1
5	3	0.54	0.75	0.39	0.46	0.57	0.74	0.00	1.00	0.00	0.00	0.68	0.53	0.74	0.42	0.47	0.93	0	1	0	0	17
5	4	0.58	0.71	0.50	0.50	0.58	0.72	0.24	0.89	0.42	0.31	0.70	0.44	0.80	0.44	0.44	0.93	0	1	0	0	13
5	5	0.57	0.71	0.47	0.48	0.58	0.74	0.26	0.91	0.50	0.34	0.68	0.38	0.79	0.39	0.39	0.93	0	0.99	0	0	19
5	6	0.59	0.71	0.51	0.50	0.59	0.71	0.35	0.84	0.43	0.38	0.75	0.43	0.87	0.54	0.48	0.93	0	1	0	0	6
5	7	0.62	0.68	0.57	0.53	0.59	0.71	0.35	0.83	0.42	0.38	0.74	0.49	0.83	0.51	0.50	0.93	0	1	0	0	6
4	3	0.55	0.84	0.34	0.47	0.61	0.74	0.00	1.00	0.00	0.00	0.71	0.46	0.80	0.45	0.45	0.93	0	1	0	0	12
4	4	0.58	0.83	0.40	0.49	0.62	0.72	0.05	0.95	0.23	0.08	0.70	0.41	0.80	0.43	0.42	0.93	0	1	0	0	16
4	5	0.61	0.80	0.47	0.52	0.63	0.73	0.17	0.93	0.44	0.24	0.72	0.46	0.81	0.46	0.46	0.93	0	1	0	0	4
4	6	0.60	0.78	0.48	0.51	0.62	0.73	0.27	0.89	0.45	0.34	0.73	0.40	0.85	0.49	0.44	0.93	0	1	0	0	5
4	7	0.63	0.76	0.54	0.54	0.63	0.73	0.33	0.86	0.46	0.39	0.75	0.44	0.86	0.53	0.48	0.93	0	0.99	0	0	3
3	3	0.53	0.77	0.36	0.46	0.58	0.74	0.00	1.00	0.00	0.00	0.64	0.40	0.73	0.34	0.37	0.93	0	1	0	0	20
3	4	0.55	0.85	0.34	0.47	0.61	0.73	0.23	0.91	0.46	0.30	0.74	0.27	0.92	0.53	0.35	0.93	0	1	0	0	10
3	5	0.58	0.81	0.42	0.49	0.61	0.72	0.29	0.87	0.43	0.35	0.72	0.29	0.87	0.43	0.35	0.93	0	1	0	0	11
3	6	0.57	0.72	0.46	0.48	0.58	0.71	0.36	0.83	0.43	0.39	0.72	0.29	0.87	0.46	0.36	0.93	0	1	0	0	13
3	7	0.60	0.69	0.53	0.51	0.58	0.74	0.38	0.87	0.50	0.43	0.71	0.40	0.82	0.44	0.42	0.93	0.05	0.99	0.50	0.11	9
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Twenty random forest (RF) models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. Despite having an equally poor performance on detecting high-risk images compared to logistic

regression and SVM, the RF models showed a marginally better performance in identifying normal, lichenoid and low-risk images. The best performing RF model using this analysis approach was the 5-feature set with k=7 for measurement clustering (Table 6.16.).

Table 6.16. Test results of all random forest models developed across all combinations of feature sets and clustering for all diagnostic categories on acriflavine images

Combination		Normal					Lichenoid					Low-risk					High-risk					Model rank
Features	Clusters (k)	Acc	Sen	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	
6	3	0.56	0.52	0.60	0.47	0.50	0.67	0.44	0.75	0.38	0.41	0.66	0.29	0.79	0.34	0.32	0.91	0	0.98	0	0	19
6	4	0.57	0.59	0.56	0.48	0.53	0.69	0.36	0.80	0.39	0.38	0.67	0.34	0.79	0.37	0.35	0.92	0.06	0.98	0.20	0.09	15
6	5	0.55	0.54	0.56	0.46	0.50	0.65	0.39	0.74	0.35	0.37	0.69	0.31	0.83	0.40	0.35	0.92	0.06	0.98	0.17	0.09	18
6	6	0.56	0.63	0.52	0.48	0.55	0.71	0.36	0.83	0.43	0.39	0.71	0.40	0.83	0.45	0.42	0.93	0	1.00	0	0	10
6	7	0.62	0.77	0.52	0.53	0.63	0.75	0.32	0.90	0.51	0.39	0.76	0.50	0.86	0.56	0.53	0.93	0	1.00	0	0	3
5	3	0.54	0.56	0.52	0.45	0.50	0.67	0.33	0.78	0.34	0.34	0.70	0.37	0.82	0.42	0.39	0.92	0	0.99	0	0	17
5	4	0.62	0.56	0.66	0.53	0.54	0.70	0.46	0.78	0.42	0.44	0.69	0.41	0.79	0.41	0.41	0.91	0	0.98	0	0	12
5	5	0.56	0.60	0.52	0.47	0.53	0.71	0.36	0.83	0.43	0.39	0.68	0.37	0.79	0.39	0.38	0.93	0	1.00	0	0	14
5	6	0.60	0.77	0.48	0.51	0.61	0.74	0.29	0.90	0.49	0.36	0.72	0.40	0.84	0.47	0.43	0.93	0	1.00	0	0	4
5	7	0.70	0.75	0.66	0.61	0.67	0.76	0.44	0.87	0.55	0.49	0.77	0.59	0.84	0.57	0.58	0.93	0.06	0.99	0.25	0.10	1
4	3	0.52	0.60	0.46	0.44	0.51	0.67	0.21	0.83	0.30	0.25	0.70	0.37	0.82	0.42	0.39	0.91	0.00	0.98	0.00	0.00	20
4	4	0.55	0.60	0.50	0.46	0.52	0.69	0.27	0.84	0.37	0.31	0.72	0.46	0.82	0.48	0.47	0.93	0.06	0.99	0.25	0.10	13
4	5	0.60	0.59	0.60	0.51	0.55	0.68	0.35	0.79	0.37	0.36	0.72	0.49	0.81	0.48	0.48	0.93	0	1.00	0	0	11
4	6	0.60	0.68	0.55	0.51	0.59	0.69	0.30	0.82	0.37	0.33	0.74	0.46	0.84	0.50	0.48	0.93	0	1.00	0	0	8
4	7	0.63	0.76	0.54	0.54	0.63	0.73	0.39	0.84	0.46	0.43	0.76	0.40	0.89	0.56	0.47	0.93	0.06	1.00	0.50	0.11	2

3	3	0.56	0.61	0.52	0.47	0.53	0.72	0.21	0.89	0.40	0.28	0.62	0.40	0.70	0.33	0.36	0.93	0	1.00	0	0	16
3	4	0.62	0.63	0.62	0.54	0.58	0.72	0.41	0.83	0.45	0.43	0.70	0.44	0.80	0.44	0.44	0.92	0	0.98	0	0	7
3	5	0.63	0.63	0.62	0.54	0.58	0.73	0.46	0.82	0.47	0.46	0.66	0.35	0.77	0.35	0.35	0.94	0.06	1.00	1.00	0.11	6
3	6	0.58	0.61	0.56	0.49	0.55	0.72	0.33	0.85	0.44	0.38	0.66	0.40	0.76	0.37	0.38	0.94	0.12	1.00	1.00	0.21	9
3	7	0.58	0.65	0.53	0.49	0.56	0.74	0.39	0.86	0.50	0.44	0.73	0.44	0.83	0.48	0.46	0.92	0	0.99	0	0	4
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																					

Twenty XGBoost models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. The overall performance of the XGBoost models was similar to the RF models where it performed poorly on the high-risk images while having a poor to moderate performance in identifying normal, lichenoid, and low-risk images. The best performing XGBoost model using this analysis approach was also the 5-feature set with k=7 for measurement clustering (Table 6.17.).

Table 6.17. Test results of all XGBoost models developed across all combinations of feature sets and clustering for all diagnostic categories on acriflavine images

Combination		**Normal**					**Lichenoid**					**Low-risk**					**High-risk**					**Model rank**
Features	**Clusters (k)**	**Acc**	**Sen**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	
6	3	0.53	0.56	0.51	0.44	0.49	0.67	0.35	0.78	0.35	0.35	0.68	0.27	0.83	0.35	0.30	0.91	0.00	0.97	0.00	0.00	20
6	4	0.58	0.59	0.56	0.49	0.54	0.69	0.36	0.81	0.39	0.38	0.69	0.37	0.80	0.40	0.38	0.93	0.06	0.99	0.25	0.10	14
6	5	0.56	0.56	0.57	0.48	0.51	0.68	0.35	0.79	0.37	0.36	0.69	0.40	0.79	0.41	0.40	0.93	0.06	0.99	0.25	0.10	16
6	6	0.62	0.69	0.56	0.53	0.60	0.74	0.44	0.84	0.49	0.46	0.73	0.41	0.85	0.49	0.45	0.93	0.00	0.99	0.00	0.00	2
6	7	0.60	0.73	0.50	0.51	0.60	0.70	0.27	0.85	0.39	0.32	0.77	0.50	0.87	0.58	0.54	0.93	0.00	1.00	0.00	0.00	4
5	3	0.54	0.56	0.53	0.45	0.50	0.67	0.36	0.78	0.36	0.36	0.69	0.32	0.82	0.39	0.35	0.93	0.06	0.99	0.25	0.10	18

5	4	0.59	0.59	0.58	0.50	0.54	0.68	0.35	0.80	0.37	0.36	0.71	0.43	0.81	0.45	0.44	0.92	0.00	0.98	0.00	0.00	12
5	5	0.59	0.60	0.58	0.50	0.55	0.71	0.44	0.80	0.43	0.44	0.68	0.35	0.80	0.39	0.37	0.93	0.00	1.00	0.00	0.00	11
5	6	0.58	0.63	0.54	0.49	0.55	0.70	0.35	0.83	0.41	0.38	0.71	0.40	0.83	0.45	0.42	0.94	0.18	1.00	0.75	0.29	9
5	7	0.69	0.72	0.66	0.60	0.65	0.75	0.39	0.87	0.51	0.44	0.73	0.54	0.80	0.49	0.52	0.93	0.06	0.99	0.25	0.10	1
4	3	0.56	0.72	0.44	0.48	0.57	0.67	0.15	0.85	0.26	0.19	0.70	0.37	0.83	0.43	0.40	0.93	0.00	1.00	0.00	0.00	15
4	4	0.53	0.53	0.52	0.44	0.48	0.70	0.35	0.82	0.40	0.37	0.67	0.35	0.79	0.38	0.36	0.92	0.12	0.98	0.25	0.16	19
4	5	0.54	0.74	0.40	0.46	0.57	0.71	0.21	0.89	0.39	0.28	0.72	0.35	0.85	0.46	0.40	0.93	0.00	1.00	0.00	0.00	13
4	6	0.61	0.64	0.58	0.52	0.57	0.72	0.41	0.83	0.45	0.43	0.72	0.43	0.82	0.46	0.44	0.92	0.00	0.99	0.00	0.00	3
4	7	0.62	0.64	0.60	0.53	0.58	0.70	0.47	0.78	0.42	0.44	0.73	0.34	0.87	0.48	0.40	0.91	0.00	0.97	0.00	0.00	5
3	3	0.56	0.57	0.56	0.47	0.52	0.71	0.26	0.87	0.41	0.32	0.63	0.40	0.71	0.33	0.36	0.92	0.06	0.98	0.17	0.09	17
3	4	0.59	0.60	0.58	0.50	0.55	0.71	0.36	0.83	0.43	0.39	0.69	0.44	0.78	0.42	0.43	0.92	0.00	0.99	0.00	0.00	8
3	5	0.60	0.61	0.60	0.52	0.56	0.74	0.44	0.84	0.48	0.46	0.68	0.37	0.79	0.39	0.38	0.92	0.06	0.98	0.17	0.09	7
3	6	0.57	0.54	0.60	0.48	0.51	0.72	0.46	0.81	0.46	0.46	0.69	0.43	0.79	0.42	0.42	0.93	0.12	0.99	0.50	0.19	10
3	7	0.58	0.59	0.56	0.49	0.54	0.72	0.36	0.84	0.44	0.40	0.75	0.54	0.83	0.53	0.54	0.92	0.00	0.99	0.00	0.00	6
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Comparing the best ML models from approach 2

The Random Forest (RF) model was the best performer in this approach, outperforming the other models across most metrics and diagnostic categories. For the no dysplasia category, the RF model achieved an accuracy of 0.70, sensitivity of 0.75, and F1-score of 0.67, showing an improvement over the Approach 1 results. The lichenoid category also saw improved performance, with the RF model reaching an accuracy of 0.763, sensitivity of 0.44, and F1-score of 0.49 (Table 6.18. & 6.19.).

Table 6.18. Approach 2 test results for the best ranking ML model for each ML model type with respect to each diagnostic category (1 vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	0.55	0.67	0.70	0.69
	sensitivity	0.81	0.79	0.75	0.72
	specificity	0.36	0.58	0.66	0.66
	precision	0.47	0.57	0.61	0.60
	f1score	0.60	0.66	0.67	0.65
lichenoid	accuracy	0.74	0.72	0.76	0.75
	sensitivity	0.18	0.32	0.44	0.39
	specificity	0.93	0.86	0.87	0.87
	precision	0.46	0.45	0.55	0.51
	f1score	0.26	0.37	0.49	0.44
low-risk	accuracy	0.70	0.75	0.77	0.73
	sensitivity	0.29	0.49	0.59	0.54
	specificity	0.85	0.84	0.84	0.80
	precision	0.41	0.52	0.57	0.49
	f1score	0.34	0.50	0.58	0.52
high-risk	accuracy	0.93	0.93	0.93	0.93
	sensitivity	0	0	0.06	0.06
	specificity	1	1	0.99	0.99
	precision	0	0	0.25	0.25
	f1score	0	0	0.10	0.10

The low-risk category had the highest performance, with the RF model achieving an accuracy of 0.77, sensitivity of 0.59, and F1-score of 0.58. Similar to Approach 1, the high-risk category remained the most challenging, with the RF model unable to achieve high sensitivity (0.06) and precision (0.25), resulting in a low F1-score of 0.10 (Table 6.18. & 6.19.).

Table 6.19. Approach 2 ranks for the best ranking ML model for each ML model type with respect to each diagnostic category (1 vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	4	3	1	2
	sensitivity	1	2	3	4
	specificity	4	3	1	1
	precision	4	3	1	2
	f1score	4	2	1	3
lichenoid	accuracy	3	4	1	2
	sensitivity	4	3	1	2
	specificity	1	4	2	3
	precision	3	4	1	2
	f1score	4	3	1	2
low-risk	accuracy	4	2	1	3
	sensitivity	4	3	1	2
	specificity	1	2	2	4
	precision	4	2	1	3
	f1score	4	3	1	2
high-risk	accuracy	1	1	3	3
	sensitivity	3	3	1	1
	specificity	1	1	3	3
	precision	3	3	1	1
	f1score	3	3	1	1
Aggregate rank (sum)		60	54	28	46

Final rank	4	3	1	2

6.5.1.3. Approach 3: Spatial distribution

All mean and standard deviations for all distances between the centroids of nuclei in each image were calculated and assessed using summary statistics (Figure 6.10.).

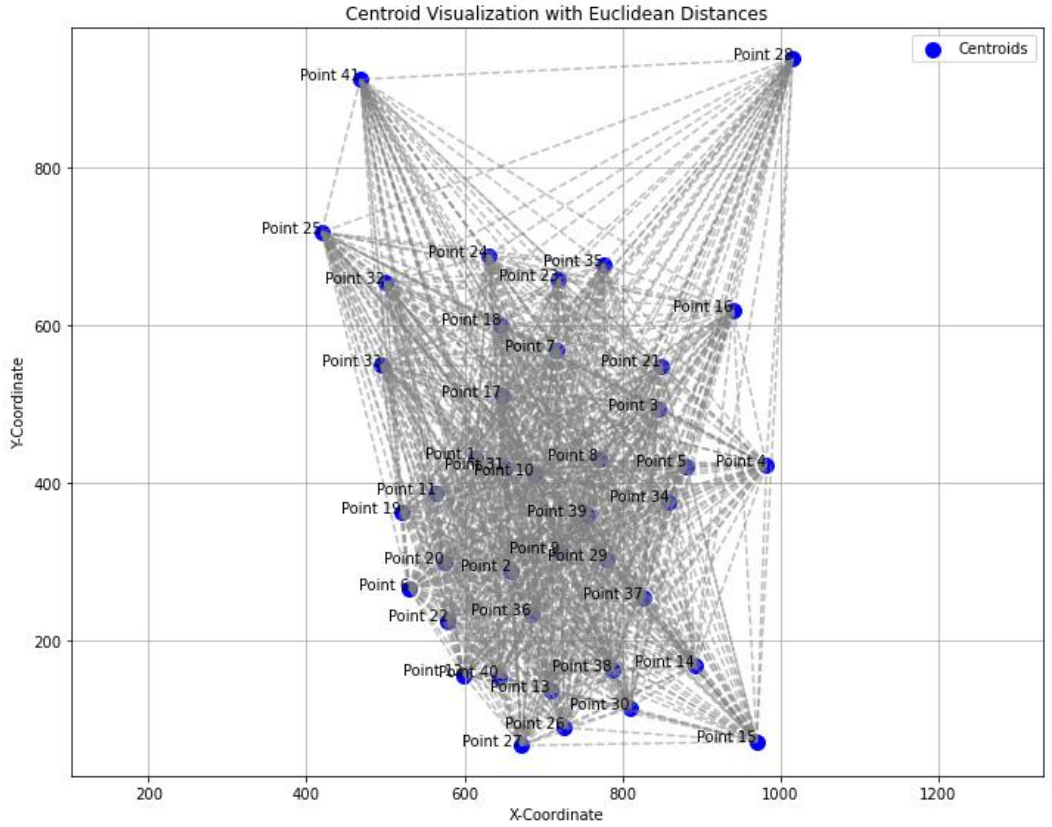


Figure 6.10. Graphical representation of nucleus distance measurements for an example acriflavine image

The no dysplasia images showed the highest mean and standard deviation values on average compared to all other diagnostic categories while the high-risk images showed the lowest. This indicates the no dysplasia images tended to have more spaced-out nuclei with a larger variability in their inter-nuclei distance. High-risk images on the other hand tended to have nuclei closer to each other with comparatively lesser variance in their distance measurements (Table 6.20.).

Table 6.20. Summary of means and standard deviation of distances between all acriflavine nuclei per image

		Normal	**Lichenoid**	**Low-risk**	**High-risk**	**P value for ANOVA**
mean distance	mean	481.34	478.90	478.91	442.32	**<0.001**
	median	489.85	483.20	483.74	461.06	
	s.d.	68.56	51.48	69.67	97.95	
s.d. of distance	mean	233.01	234.35	234.05	218.20	**0.004**
	median	240.22	237.69	237.14	226.18	
	s.d.	40.03	27.31	34.56	50.16	

The one-way ANOVA further sheds light on the presence of significant differences in the mean and standard deviation of nuclei distances between the groups (Table 6.20.). The Tukey's pairwise comparison of the diagnostic groups further shows that for both measurement metrics the measurements in the high-risk images were statistically significantly different compared to all other groups (Table 6.21.).

Table 6.21. Results of Tukey's pairwise comparison post-hoc test of categories mean and standard deviation of distance measurements between nuclei across all diagnostic categories in acriflavine images

	group 1	**group 2**	**mean difference**	**lower 95% confidence interval**	**upper 95% confidence interval**	**p value**
mean distance	high-risk	lichenoid	36.581	14.16	59.00	**<0.001**
	high-risk	low-risk	36.5904	15.20	57.99	**<0.001**
	high-risk	normal	39.023	17.95	60.10	**<0.001**
	lichenoid	low-risk	0.0094	-13.88	13.90	1
	lichenoid	normal	2.442	-10.95	15.83	0.966
	low-risk	normal	2.4326	-9.16	14.02	0.949
s.d. of distance	high-risk	lichenoid	16.151	4.053	28.249	**0.003**
	high-risk	low-risk	15.855	4.312	27.399	**0.002**
	high-risk	normal	14.809	3.439	26.179	**0.005**
	lichenoid	low-risk	-0.296	-7.791	7.200	1
	lichenoid	normal	-1.342	-8.567	5.883	0.964
	low-risk	normal	-1.046	-7.299	5.206	0.973

In terms of ML models developed on these two distance measurements the logistic regression (LR) model was the best performer in this approach, achieving the highest overall accuracy and F1-score (Tables 6.22. & 6.23.).

Table 6.22. Approach 3 test results for the 4 ML models with respect to each diagnostic category (1 vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
normal	accuracy	0.40	0.52	0.56	0.39
	sensitivity	1	0.87	0.50	0.92
	specificity	0.01	0.29	0.60	0.05
	precision	0.40	0.44	0.44	0.38
	f1score	0.57	0.59	0.47	0.54
lichenoid	accuracy	0.79	0.78	0.70	0.79
	sensitivity	0	0	0.15	0
	specificity	1	0.99	0.84	1
	precision	0	0	0.21	0
	f1score	0	0	0.18	0
low-risk	accuracy	0.67	0.63	0.59	0.64
	sensitivity	0	0.28	0.41	0.04
	specificity	1	0.81	0.68	0.93
	precision	0	0.41	0.38	0.20
	f1score	0	0.33	0.40	0.06
high-risk	accuracy	0.94	0.93	0.89	0.93
	sensitivity	0.06	0	0.06	0
	specificity	1	1	0.95	1
	precision	1	0	0.08	0
	f1score	0.12	0	0.07	0

For the no dysplasia images, the LR model achieved an accuracy of 0.398, sensitivity of 1.000, and F1-score of 0.566, indicating a high sensitivity but lower precision. Despite being the best performing model, it did not correctly identify a single instance of lichenoid and low-risk cases. The high-risk category remained challenging, with the LR model achieving an accuracy of 0.939, sensitivity of 0.063, and F1-score of 0.118 (Tables 6.22. & 6.23.).

Table 6.23. Approach 3 ranks for the 4 ML models with respect to each diagnostic category (1 vs all) in the acriflavine test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
normal	accuracy	3	2	1	4
	sensitivity	1	3	4	2
	specificity	4	2	1	3
	precision	3	2	1	4
	f1score	2	1	4	3
lichenoid	accuracy	1	3	4	1
	sensitivity	2	2	1	2
	specificity	1	3	4	1
	precision	2	2	1	2
	f1score	2	2	1	2
low-risk	accuracy	1	3	4	2
	sensitivity	4	2	1	3
	specificity	1	3	4	2
	precision	4	1	2	3
	f1score	4	2	1	3
high-risk	accuracy	1	2	4	2
	sensitivity	1	3	1	3
	specificity	1	1	4	1
	precision	1	3	2	3
	f1score	1	3	2	3
Aggregate rank (sum)		40	45	47	49
Final rank		1	2	3	4

6.5.1.4. Comparison of best feature extraction ML model results for acriflavine

All 3 approaches for feature extraction on the acriflavine dataset followed by ML diagnostic triage produced varying performance (Table 6.24.).

Table 6.24. Test results of the best ranked ML models based on acriflavine segmentation data for all 3 analysis approaches

Diagnostic category	Metric	Best approach 1 model (SVM)	Best approach 2 model (RF)	Best approach 3 model (LR)
no dysplasia	accuracy	0.64	0.70	0.40
	sensitivity	0.59	0.75	1
	specificity	0.67	0.66	0.01
	precision	0.52	0.61	0.40
	f1score	0.55	0.67	0.57
lichenoid	accuracy	0.70	0.76	0.79
	sensitivity	0.43	0.44	0
	specificity	0.81	0.87	1
	precision	0.49	0.55	0
	f1score	0.46	0.49	0
low-risk	accuracy	0.75	0.77	0.67
	sensitivity	0.55	0.59	0
	specificity	0.83	0.84	1
	precision	0.55	0.57	0
	f1score	0.55	0.58	0
high-risk	accuracy	0.94	0.93	0.94
	sensitivity	0.20	0.06	0.06
	specificity	0.98	0.99	1
	precision	0.33	0.25	1
	f1score	0.25	0.10	0.12

Upon ranking the best model from each approach, the Approach 2 model which was a random forest model performed the best (Table 6.25.). This ML model had a moderately high sensitivity (0.75) at detecting ‘No dysplasia’ cases with a moderate specificity (0.66). It performed worse on detecting ‘Lichenoid’ cases with a sensitivity of 0.44. The Approach 2 model detected only just over half of the ‘Low-risk’ cases (sensitivity = 0.59) with slightly above half of its ‘Low-risk’ predictions being correct (precision = 0.57). However, the model performance for detecting ‘High-risk’ cases was exceptionally poor with it only correctly detecting about 6% of the cases (sensitivity = 0.06) (Table 24). Clinically this model would be unfit for diagnostic triage as it is not substantially superior to a randomly guessing model (AUC 0.5) for non-dysplastic tissue and low-grade OED, with an inability to detect high-grade OED and OSCC.

Table 6.25. Ranks of the best ranked ML models based on acriflavine segmentation data for all 3 analysis approaches

Diagnostic category	Metric	Best approach 1 model (SVM)	Best approach 2 model (RF)	Best approach 3 model (LR)
no dysplasia	accuracy	2	1	3
	sensitivity	3	2	1
	specificity	1	2	3
	precision	2	1	3
	f1score	3	1	2
lichenoid	accuracy	3	2	1
	sensitivity	2	1	3
	specificity	3	2	1
	precision	2	1	3
	f1score	2	1	3
low-risk	accuracy	2	1	3
	sensitivity	2	1	3
	specificity	3	2	1
	precision	2	1	3
	f1score	2	1	3
high-risk	accuracy	1	3	2
	sensitivity	1	3	2
	specificity	3	2	1
	precision	2	3	1
	f1score	1	3	2
Aggregate rank		42	34	44
Final rank		2	1	3

6.5.2. Fluorescein ML performance results

All measurement data across the 6 measurements of mean pixel intensity, standard deviation of pixel intensity, mean surface area, integrated density, circularity and aspect ratio of all 10,143 fluorescein-stained nuclei were standardised using the z-score method to have a mean of 0 and a standard deviation of 1. There were 4 ML models developed for each of the 4 model types across all approaches 1 & 3, with 80 models being developed in approach 2 resulting in a total of 96 ML models being developed.

6.5.2.1. Approach 1: Direct classification of measurement means for size, shape and brightness.

Across all 4 models developed on this dataset, the random forest (RF) model was also the best performer in this approach, outperforming the other models across most metrics and diagnostic categories in addition to the best acriflavine approach 1 ML model (Table 6.26.).

Table 6.26. Approach 1 test results for all 4 ML models with respect to each diagnostic category (1 vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
normal	accuracy	0.684	0.772	0.702	0.702
	sensitivity	0.40	0.47	0.67	0.40
	specificity	0.79	0.88	0.71	0.81
	precision	0.40	0.58	0.46	0.43
	f1score	0.40	0.52	0.54	0.41
lichenoid	accuracy	0.684	0.596	0.632	0.632
	sensitivity	0.62	0.62	0.50	0.58
	specificity	0.74	0.58	0.74	0.68
	precision	0.67	0.55	0.62	0.60
	f1score	0.64	0.58	0.55	0.59
low-risk	accuracy	0.825	0.807	0.842	0.789
	sensitivity	0.43	0.29	0.43	0.43
	specificity	0.88	0.88	0.90	0.84
	precision	0.33	0.25	0.38	0.27
	f1score	0.38	0.27	0.40	0.33
high-risk	accuracy	0.86	0.912	0.912	0.895
	sensitivity	0.56	0.67	0.56	0.56
	specificity	0.92	0.96	0.98	0.96
	precision	0.56	0.75	0.83	0.71
	f1score	0.56	0.71	0.67	0.63

The RF model had moderate to high accuracy for no dysplasia (70.2%), lichenoid (63.2%), and low-risk (84.2%) images with it being most accurate with high-risk images (91.2%). This is reflected in the moderate to high sensitivity, specificity, and precision scores with the model performing best on high-risk images (F1 score = 0.667) (Tables 6.26. & 6.27.).

Table 6.27. Approach 1 ranks for all 4 ML models with respect to each diagnostic category (1vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	4	1	2	2
	sensitivity	3	2	1	3
	specificity	3	1	4	2
	precision	4	1	2	3
	f1score	4	2	1	3
lichenoid	accuracy	1	4	2	2
	sensitivity	1	1	4	3
	specificity	1	4	1	3
	precision	1	4	2	3
	f1score	1	3	4	2
low-risk	accuracy	2	3	1	4
	sensitivity	1	4	1	1
	specificity	2	2	1	4
	precision	2	4	1	3
	f1score	2	4	1	3
high-risk	accuracy	4	1	1	3
	sensitivity	2	1	2	2
	specificity	4	2	1	2
	precision	4	2	1	3
	f1score	4	1	2	3
Aggregate rank (sum)		50	47	35	54
Final rank		3	2	1	4

6.5.2.2. Approach 2: Clustering and feature selection for size, shape and brightness

The within cluster sum of squares (WCSS) representing the cohesiveness of the grouping of nuclei measurements gets better as the number of clusters (k) is increased. This indicates the clusters of nuclei measurement data become more grouped up as k is increased, however that also increases the complexity of the model which can significantly increase computational costs and time (Figure 6.11.).

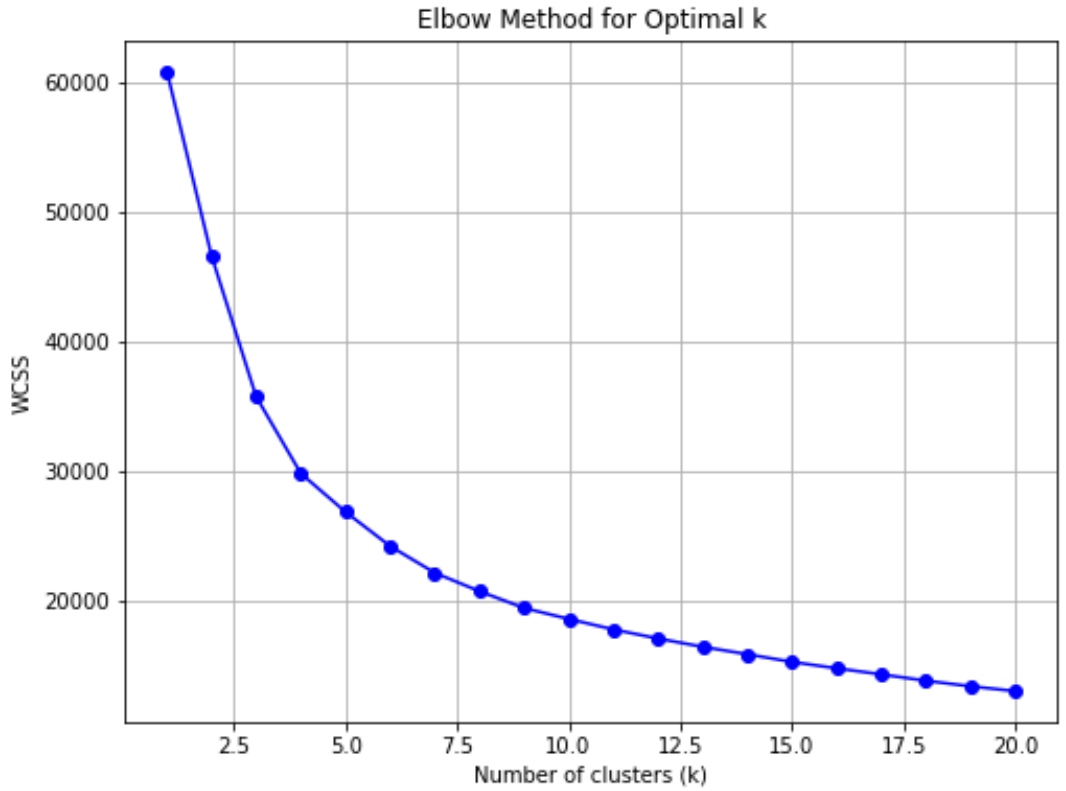


Figure 6.11. Elbow plot of WCSS vs number of clusters for fluorescein nuclei measurements

The 'elbow' in the plot depicted in Figure 11 appears to be around k=4, k=5, k=6 and k=7 which would make these the ideal candidates to test in the fluorescein approach 2 models.

Based on the spearman's correlation matrix (Figure 6.12.) mean pixel intensity and area were highly correlated with integrated density. Circularity and aspect ratio were also highly correlated with each other.

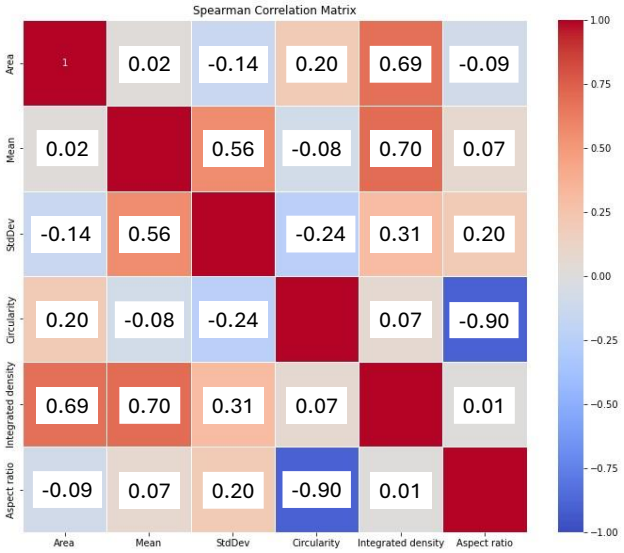


Figure 6.12. Spearman's correlation matrix for fluorescein nuclei measurements

Three features were eliminated in a step-wise manner when comparing pairs of features with the highest correlation (Figure 6.13.). At each step a new feature set was created beginning with the original 6-feature set. The first pair to have the highest absolute correlation were circularity & aspect ratio (0.90). The sum of absolute correlation of both these metrics against the rest:

- Circularity = 0.07 + 0.24 + 0.08 + 0.20 = 0.59
- Aspect ratio = 0.01 + 0.20 + 0.07 + 0.09 = 0.37

Since the absolute sum of correlation for aspect ratio against all other metrics is lower (0.37), this indicates it is less correlated with the other metrics overall compared to circularity. Thus, circularity was eliminated and aspect ratio retained. The 5-feature set comprised of the remaining features of mean pixel intensity, area, standard deviation of pixel intensity, integrated density, and aspect ratio (Figure 6.13.B).

The next pair with the highest absolute correlation was mean pixel intensity and integrated density (0.70). The sum of absolute correlation of both these metrics against the rest:

- Mean = 0.07 + 0.56 + 0.02 = 0.65
- Integrated density = 0.01 + 0.31 + 0.69 = 1.01

Since the absolute sum of correlation for mean intensity against all other metrics is lower (0.65), this indicates it is less correlated with the other metrics overall compared to integrated density. Thus, integrated density was eliminated and mean intensity retained. The 4-feature

set comprised of the remaining features of mean pixel intensity, area, standard deviation of pixel intensity, and aspect ratio (Figure b.13.C).

The next pair with the largest correlation was mean intensity and standard deviation of intensity (0.56). The sum of absolute correlation of both these metrics against the rest:

- Mean = 0.07 + 0.02 = 0.09
- S.d. = 0.20 + 0.14 = 0.34

Since the absolute sum of correlation for mean intensity against all other metrics is lower (0.104), this indicates it is less correlated with the other metrics overall compared to standard deviation of intensity. Thus, standard deviation of intensity was eliminated and mean intensity retained. The 3-feature set comprised of the remaining features of mean pixel intensity, area, and aspect ratio (Figure 6.13.D).

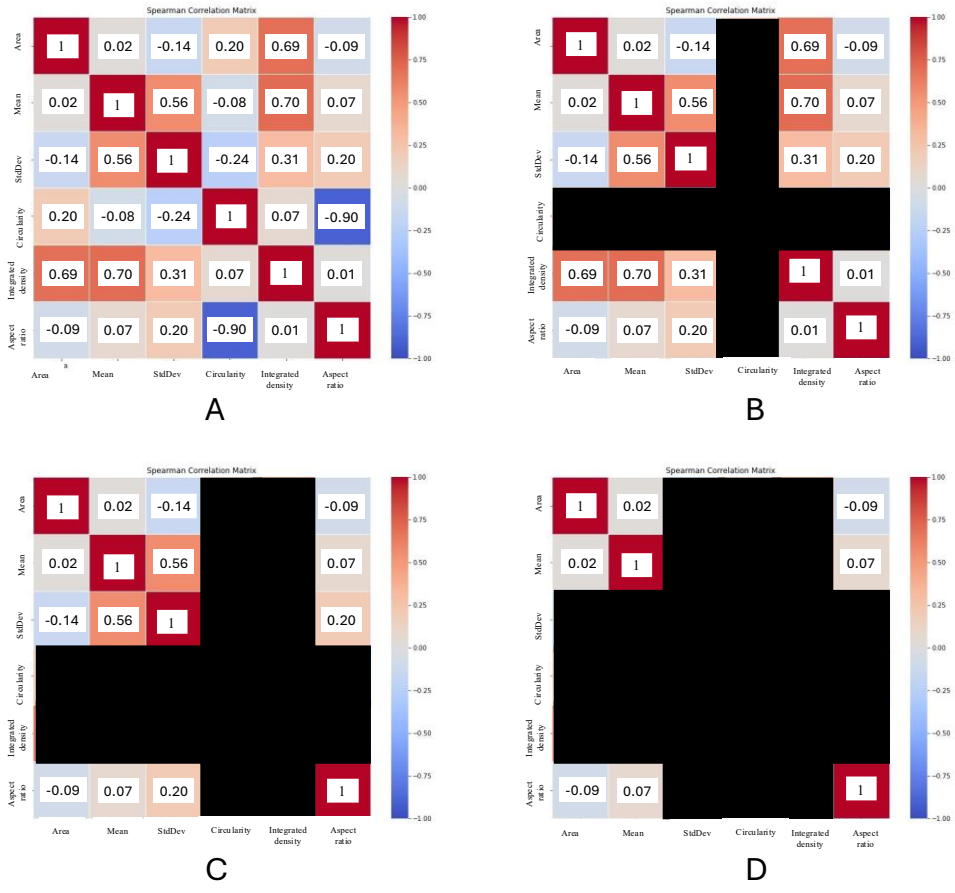


Figure 6.13. Spearman's correlation matrix of the 6 measurement features of the nuclei in fluorescein images. The four panels show which features were eliminated at each step of the feature selection. A) original 6-feature set, B) 5-feature set, C) 4-feature set, D) 3-feature set

ML test results for approach 2

Twenty logistic regression models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. The models across all combinations had very poor results in correctly identifying high-risk images. They also performed poorly with normal, lichenoid and low-risk images. Upon ranking them the best performing model was the one developed on the 6-feature set with k=4 for measurement clustering (Table 6.28.).

Table 6.28. Test results of all logistic regression models developed across all combinations of feature sets and clustering for all diagnostic categories on fluorescein images

Combination		Normal					Lichenoid					Low-risk					High-risk					Model rank
Features	Clusters (k)	Acc	Sen	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	
6	3	0.47	0.53	0.44	0.27	0.36	0.57	0.53	0.58	0.35	0.42	0.75	0	1	0	0	0.83	0	1	0	0	20
6	4	0.66	0.38	0.76	0.38	0.38	0.62	0.58	0.64	0.40	0.48	0.68	0.45	0.76	0.39	0.42	0.83	0	1	0	0	1
6	5	0.62	0.24	0.77	0.29	0.26	0.64	0.58	0.66	0.42	0.49	0.63	0.42	0.70	0.33	0.37	0.83	0.10	0.98	0.50	0.16	9
6	6	0.61	0.38	0.69	0.33	0.35	0.57	0.50	0.59	0.34	0.40	0.67	0.32	0.79	0.35	0.33	0.83	0	1	0	0	18
6	7	0.63	0.38	0.73	0.35	0.37	0.57	0.44	0.63	0.33	0.38	0.64	0.39	0.73	0.32	0.35	0.83	0	1	0	0	17
5	3	0.62	0.53	0.66	0.38	0.44	0.62	0.61	0.62	0.40	0.48	0.74	0.29	0.89	0.47	0.36	0.83	0	1	0	0	3
5	4	0.61	0.24	0.75	0.27	0.25	0.62	0.50	0.66	0.38	0.43	0.66	0.48	0.73	0.38	0.42	0.82	0.10	0.97	0.40	0.15	15
5	5	0.62	0.53	0.65	0.37	0.43	0.60	0.44	0.66	0.36	0.40	0.70	0.36	0.81	0.39	0.37	0.83	0	1	0	0	12
5	6	0.62	0.50	0.67	0.37	0.43	0.62	0.47	0.67	0.38	0.42	0.69	0.32	0.81	0.37	0.35	0.83	0.10	0.98	0.50	0.16	5
5	7	0.66	0.35	0.77	0.38	0.36	0.59	0.50	0.63	0.36	0.42	0.67	0.45	0.75	0.38	0.41	0.84	0.10	0.99	0.67	0.17	6
4	3	0.57	0.47	0.61	0.32	0.38	0.62	0.47	0.67	0.38	0.42	0.72	0.39	0.84	0.44	0.41	0.83	0	1	0	0	8
4	4	0.58	0.21	0.73	0.23	0.22	0.66	0.56	0.71	0.44	0.49	0.68	0.52	0.74	0.40	0.45	0.81	0.10	0.96	0.33	0.15	10
4	5	0.64	0.41	0.73	0.37	0.39	0.63	0.42	0.72	0.39	0.40	0.66	0.42	0.74	0.35	0.38	0.80	0.10	0.94	0.25	0.14	14

4	6	0.64	0.50	0.69	0.39	0.44	0.65	0.47	0.72	0.42	0.44	0.66	0.36	0.77	0.34	0.35	0.82	0.10	0.97	0.40	0.15	7
4	7	0.62	0.44	0.68	0.35	0.39	0.61	0.42	0.69	0.36	0.39	0.67	0.45	0.75	0.38	0.41	0.83	0	1	0	0	13
3	3	0.52	0.44	0.55	0.27	0.34	0.67	0.56	0.72	0.46	0.50	0.69	0.26	0.84	0.35	0.30	0.83	0	1	0	0	10
3	4	0.58	0.50	0.61	0.33	0.40	0.66	0.53	0.71	0.43	0.48	0.72	0.32	0.86	0.44	0.37	0.83	0.10	0.98	0.50	0.16	2
3	5	0.59	0.38	0.67	0.31	0.34	0.62	0.56	0.65	0.40	0.47	0.71	0.36	0.84	0.42	0.39	0.83	0.10	0.98	0.50	0.16	4
3	6	0.56	0.27	0.67	0.24	0.25	0.61	0.56	0.63	0.39	0.46	0.65	0.32	0.76	0.31	0.32	0.83	0	1	0	0	19
3	7	0.61	0.38	0.69	0.33	0.35	0.59	0.56	0.61	0.37	0.44	0.69	0.32	0.81	0.37	0.35	0.82	0	0.99	0	0	16
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Twenty SVM models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. The models across all combinations had poor to moderate performance across all diagnostic categories. Upon ranking them the best performing model was the one developed on the 5-feature set with k=7 for measurement clustering (Table 6.29.).

Table 6.29. Test results of all SVM models developed across all combinations of feature sets and clustering for all diagnostic categories on fluorescein images

Combination		**Normal**					**Lichenoid**					**Low-risk**					**High-risk**					**Model rank**
Features	**Clusters (k)**	**Acc**	**Sen**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	
6	3	0.47	0.68	0.39	0.30	0.41	0.56	0.36	0.64	0.30	0.33	0.74	0	0.99	0	0	0.83	0	1	0	0	18
6	4	0.57	0.35	0.66	0.29	0.32	0.62	0.36	0.72	0.35	0.36	0.65	0.48	0.70	0.36	0.41	0.84	0.05	1	1	0.09	16
6	5	0.68	0.29	0.83	0.40	0.34	0.67	0.42	0.78	0.44	0.43	0.65	0.61	0.66	0.38	0.47	0.77	0.14	0.90	0.23	0.18	8
6	6	0.58	0.47	0.63	0.33	0.39	0.57	0.33	0.66	0.29	0.31	0.68	0.26	0.82	0.33	0.29	0.81	0.14	0.95	0.38	0.21	18
6	7	0.63	0.53	0.67	0.38	0.44	0.63	0.39	0.73	0.38	0.38	0.68	0.36	0.79	0.37	0.36	0.81	0.14	0.95	0.38	0.21	10

5	3	0.66	0.41	0.75	0.39	0.40	0.67	0.53	0.73	0.45	0.49	0.67	0.29	0.80	0.33	0.31	0.79	0.29	0.89	0.35	0.32	7
5	4	0.46	0.47	0.46	0.25	0.33	0.62	0.36	0.73	0.36	0.36	0.73	0.32	0.87	0.46	0.38	0.83	0	1	0	0	14
5	5	0.62	0.47	0.68	0.36	0.41	0.62	0.44	0.69	0.37	0.41	0.66	0.39	0.75	0.34	0.36	0.83	0	1	0	0	13
5	6	0.63	0.32	0.75	0.33	0.33	0.65	0.44	0.73	0.41	0.43	0.66	0.32	0.78	0.33	0.33	0.81	0.43	0.89	0.45	0.44	9
5	7	0.71	0.44	0.82	0.48	0.46	0.63	0.42	0.72	0.39	0.40	0.68	0.52	0.74	0.40	0.45	0.84	0.33	0.95	0.58	0.42	1
4	3	0.51	0.24	0.61	0.19	0.21	0.62	0.39	0.71	0.36	0.37	0.55	0.23	0.66	0.18	0.20	0.84	0.10	0.99	0.67	0.17	20
4	4	0.57	0.24	0.71	0.24	0.24	0.61	0.39	0.70	0.35	0.37	0.69	0.48	0.76	0.41	0.44	0.77	0.10	0.91	0.18	0.13	17
4	5	0.66	0.32	0.80	0.38	0.35	0.64	0.47	0.71	0.41	0.44	0.67	0.42	0.76	0.37	0.39	0.81	0.33	0.91	0.44	0.38	5
4	6	0.67	0.44	0.76	0.42	0.43	0.62	0.42	0.71	0.38	0.40	0.68	0.48	0.75	0.40	0.44	0.86	0.29	0.98	0.75	0.41	2
4	7	0.58	0.38	0.66	0.30	0.34	0.62	0.42	0.71	0.38	0.40	0.74	0.42	0.85	0.48	0.45	0.86	0.38	0.96	0.67	0.49	4
3	3	0.44	0.65	0.36	0.28	0.39	0.61	0.44	0.67	0.36	0.40	0.75	0	1	0	0	0.83	0	1	0	0	15
3	4	0.66	0.41	0.75	0.39	0.40	0.64	0.36	0.76	0.38	0.37	0.72	0.48	0.80	0.46	0.47	0.77	0.29	0.87	0.32	0.30	6
3	5	0.62	0.29	0.74	0.30	0.30	0.54	0.50	0.56	0.32	0.39	0.71	0.32	0.84	0.40	0.36	0.86	0.29	0.98	0.75	0.41	12
3	6	0.60	0.50	0.64	0.35	0.41	0.64	0.39	0.74	0.39	0.39	0.75	0.45	0.86	0.52	0.48	0.83	0.24	0.95	0.50	0.32	3
3	7	0.58	0.41	0.65	0.31	0.35	0.71	0.39	0.84	0.50	0.44	0.69	0.32	0.81	0.37	0.35	0.76	0.33	0.85	0.32	0.33	11
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Twenty RF models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. These models performed better than the logistic regression and SVM models across all combinations with moderate classification performance across all diagnostic categories. Upon ranking them the best performing model was the one developed on the 4-feature set with k=6 for measurement clustering (Table 6.30.).

Table 6.30. Test results of all RF models developed across all combinations of feature sets and clustering for all diagnostic categories on fluorescein images

Combination		Normal					Lichenoid					Low-risk					High-risk					Model rank
Features	Clusters (k)	Acc	Sen	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	
6	3	0.64	0.35	0.75	0.35	0.35	0.62	0.39	0.72	0.37	0.38	0.71	0.39	0.81	0.41	0.40	0.77	0.33	0.86	0.33	0.33	20
6	4	0.62	0.24	0.77	0.29	0.26	0.60	0.39	0.69	0.34	0.36	0.73	0.55	0.79	0.47	0.51	0.84	0.43	0.92	0.53	0.47	15
6	5	0.66	0.32	0.78	0.37	0.34	0.64	0.44	0.72	0.40	0.42	0.80	0.61	0.86	0.59	0.60	0.81	0.43	0.89	0.45	0.44	6
6	6	0.60	0.47	0.65	0.34	0.40	0.65	0.36	0.77	0.39	0.38	0.68	0.26	0.82	0.33	0.29	0.83	0.43	0.91	0.50	0.46	17
6	7	0.62	0.41	0.71	0.35	0.38	0.72	0.53	0.80	0.53	0.53	0.72	0.42	0.82	0.45	0.43	0.79	0.29	0.89	0.35	0.32	12
5	3	0.60	0.59	0.60	0.36	0.45	0.68	0.44	0.78	0.46	0.45	0.69	0.26	0.84	0.35	0.30	0.85	0.29	0.97	0.67	0.40	9
5	4	0.62	0.29	0.74	0.30	0.30	0.62	0.44	0.70	0.38	0.41	0.71	0.39	0.81	0.41	0.40	0.81	0.38	0.90	0.44	0.41	18
5	5	0.66	0.47	0.73	0.40	0.43	0.62	0.39	0.71	0.36	0.37	0.72	0.39	0.84	0.44	0.41	0.86	0.48	0.94	0.63	0.54	8
5	6	0.60	0.35	0.69	0.31	0.33	0.69	0.39	0.81	0.47	0.42	0.66	0.42	0.74	0.35	0.38	0.83	0.38	0.92	0.50	0.43	16
5	7	0.69	0.56	0.74	0.45	0.50	0.74	0.58	0.80	0.55	0.57	0.74	0.42	0.85	0.48	0.45	0.85	0.43	0.94	0.60	0.50	2
4	3	0.66	0.50	0.72	0.41	0.45	0.63	0.44	0.71	0.39	0.42	0.70	0.39	0.80	0.40	0.39	0.82	0.19	0.95	0.44	0.27	14
4	4	0.67	0.35	0.80	0.40	0.38	0.68	0.36	0.81	0.45	0.40	0.69	0.61	0.71	0.42	0.50	0.81	0.38	0.90	0.44	0.41	11
4	5	0.64	0.38	0.74	0.36	0.37	0.66	0.44	0.76	0.43	0.44	0.69	0.42	0.78	0.39	0.41	0.84	0.43	0.93	0.56	0.49	10
4	6	0.65	0.56	0.68	0.40	0.47	0.73	0.50	0.83	0.55	0.52	0.75	0.42	0.86	0.50	0.46	0.89	0.57	0.96	0.75	0.65	1
4	7	0.54	0.32	0.63	0.25	0.28	0.66	0.42	0.76	0.42	0.42	0.75	0.45	0.85	0.50	0.48	0.86	0.43	0.95	0.64	0.51	7
3	3	0.65	0.50	0.71	0.40	0.44	0.62	0.36	0.73	0.36	0.36	0.66	0.19	0.82	0.27	0.23	0.80	0.43	0.88	0.43	0.43	19
3	4	0.66	0.27	0.82	0.36	0.31	0.62	0.42	0.71	0.38	0.40	0.79	0.68	0.82	0.57	0.62	0.80	0.38	0.88	0.40	0.39	13

3	5	0.71	0.35	0.85	0.48	0.41	0.67	0.58	0.71	0.46	0.51	0.70	0.45	0.78	0.41	0.43	0.87	0.52	0.94	0.65	0.58	4
3	6	0.67	0.41	0.77	0.41	0.41	0.73	0.56	0.80	0.54	0.55	0.75	0.55	0.81	0.50	0.52	0.84	0.43	0.92	0.53	0.47	3
3	7	0.70	0.47	0.78	0.46	0.46	0.68	0.42	0.79	0.46	0.44	0.71	0.45	0.80	0.44	0.44	0.81	0.48	0.88	0.46	0.47	5
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Twenty XGBoost models were trained and tested across all combinations of clusters selected based on the elbow plot and feature sets as determined by the feature selection process. These models performed similar to the logistic regression and SVM models and worse than the RF models across all combinations with moderate to poor classification performance across all diagnostic categories. Upon ranking them the best performing model was the one developed on the 6-feature set with k=5 for measurement clustering (Table 6.31.).

Table 6.31. Test results of all XGBoost models developed across all combinations of feature sets and clustering for all diagnostic categories on fluorescein images

Combination		**Normal**					**Lichenoid**					**Low-risk**					**High-risk**					**Model rank**
Features	**Clusters (k)**	**Acc**	**Sen**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**F1**	
6	3	0.62	0.32	0.74	0.32	0.32	0.66	0.42	0.76	0.42	0.42	0.63	0.32	0.74	0.29	0.31	0.76	0.24	0.87	0.28	0.26	20
6	4	0.62	0.29	0.74	0.30	0.30	0.60	0.42	0.67	0.35	0.38	0.74	0.45	0.84	0.48	0.47	0.84	0.43	0.92	0.53	0.47	16
6	5	0.71	0.41	0.83	0.48	0.44	0.66	0.53	0.71	0.43	0.48	0.78	0.58	0.85	0.56	0.57	0.84	0.43	0.92	0.53	0.47	1
6	6	0.62	0.38	0.71	0.33	0.36	0.61	0.39	0.70	0.35	0.37	0.67	0.26	0.81	0.32	0.29	0.81	0.38	0.90	0.44	0.41	19
6	7	0.65	0.32	0.77	0.36	0.34	0.69	0.44	0.79	0.47	0.46	0.71	0.55	0.76	0.44	0.49	0.81	0.38	0.90	0.44	0.41	12
5	3	0.64	0.53	0.68	0.39	0.45	0.66	0.53	0.71	0.43	0.48	0.66	0.23	0.80	0.28	0.25	0.84	0.19	0.97	0.57	0.29	13
5	4	0.66	0.38	0.76	0.38	0.38	0.70	0.53	0.77	0.49	0.51	0.75	0.42	0.86	0.50	0.46	0.82	0.52	0.88	0.48	0.50	3

5	5	0.62	0.47	0.67	0.36	0.41	0.62	0.33	0.74	0.35	0.34	0.73	0.42	0.84	0.46	0.44	0.85	0.43	0.94	0.60	0.50	11
5	6	0.59	0.35	0.68	0.30	0.32	0.72	0.47	0.83	0.53	0.50	0.72	0.36	0.85	0.44	0.39	0.79	0.48	0.85	0.40	0.44	14
5	7	0.62	0.35	0.73	0.33	0.34	0.71	0.56	0.77	0.50	0.53	0.73	0.48	0.81	0.47	0.48	0.86	0.43	0.95	0.64	0.51	5
4	3	0.65	0.47	0.72	0.39	0.43	0.63	0.47	0.70	0.40	0.43	0.67	0.26	0.81	0.32	0.29	0.82	0.29	0.93	0.46	0.35	17
4	4	0.66	0.35	0.78	0.39	0.37	0.68	0.42	0.79	0.46	0.44	0.72	0.65	0.75	0.47	0.54	0.82	0.33	0.92	0.47	0.39	10
4	5	0.66	0.35	0.78	0.39	0.37	0.62	0.44	0.70	0.38	0.41	0.70	0.42	0.79	0.41	0.41	0.82	0.38	0.91	0.47	0.42	15
4	6	0.66	0.47	0.74	0.41	0.44	0.65	0.50	0.71	0.42	0.46	0.70	0.39	0.80	0.40	0.39	0.89	0.43	0.99	0.90	0.58	7
4	7	0.67	0.47	0.75	0.42	0.44	0.66	0.39	0.77	0.41	0.40	0.75	0.52	0.84	0.52	0.52	0.84	0.48	0.91	0.53	0.50	6
3	3	0.62	0.47	0.67	0.36	0.41	0.67	0.42	0.78	0.44	0.43	0.70	0.26	0.85	0.36	0.30	0.77	0.33	0.86	0.33	0.33	18
3	4	0.69	0.47	0.77	0.44	0.46	0.62	0.47	0.67	0.38	0.42	0.76	0.42	0.88	0.54	0.47	0.82	0.38	0.91	0.47	0.42	8
3	5	0.71	0.38	0.84	0.48	0.43	0.66	0.53	0.72	0.44	0.48	0.73	0.52	0.80	0.47	0.49	0.84	0.48	0.92	0.56	0.51	2
3	6	0.63	0.35	0.74	0.34	0.35	0.71	0.44	0.83	0.52	0.48	0.72	0.55	0.78	0.46	0.50	0.82	0.43	0.90	0.47	0.45	9
3	7	0.62	0.59	0.63	0.38	0.46	0.70	0.50	0.78	0.49	0.49	0.75	0.36	0.89	0.52	0.42	0.85	0.33	0.96	0.64	0.44	3
Legend: Acc. = accuracy, Sens. = sensitivity, Spec. = specificity, Prec. = precision, F1 = F1 score																						

Comparing the best ML models from approach 2

The Random Forest (RF) model was the best performer in this approach, outperforming the other models across most metrics and diagnostic categories. For the no dysplasia images, the RF model achieved an accuracy of 0.648, sensitivity of 0.559, and F1-score of 0.469, showing improved performance compared to Approach 1.

Table 6.32. Approach 2 test results for the best ranking ML model for each ML model type with respect to each diagnostic category (1 vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	0.66	0.71	0.65	0.71
	sensitivity	0.38	0.44	0.56	0.41
	specificity	0.76	0.82	0.68	0.83
	precision	0.38	0.48	0.40	0.48
	f1score	0.38	0.46	0.47	0.44
lichenoid	accuracy	0.62	0.63	0.73	0.66
	sensitivity	0.58	0.42	0.50	0.53
	specificity	0.64	0.72	0.83	0.71
	precision	0.40	0.39	0.55	0.43
	f1score	0.48	0.40	0.52	0.48
low-risk	accuracy	0.68	0.68	0.75	0.78
	sensitivity	0.45	0.52	0.42	0.58
	specificity	0.76	0.74	0.86	0.85
	precision	0.39	0.40	0.50	0.56
	f1score	0.42	0.45	0.46	0.57
high-risk	accuracy	0.83	0.84	0.89	0.84
	sensitivity	0	0.33	0.57	0.43
	specificity	1	0.95	0.96	0.92
	precision	0	0.58	0.75	0.53
	f1score	0	0.42	0.65	0.47

The lichenoid images saw further improvement, with the RF model reaching an accuracy of 0.730, sensitivity of 0.500, and F1-score of 0.522 (Tables 6.32. & 6.33.). The low-risk images had the highest performance, with the RF model achieving an accuracy of 0.746, sensitivity of 0.419, and F1-score of 0.456. Similar to the acriflavine dataset, the high-risk category remained the most challenging, with the RF model unable to achieve high sensitivity (0.571) and precision (0.750), resulting in a F1-score of 0.649 (Tables 6.32. & 6.33.).

Table 6.33. Approach 2 ranks for the best ranking ML model for each ML model type with respect to each diagnostic category (1 vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	3	1	4	1
	sensitivity	4	2	1	3
	specificity	3	2	4	1
	precision	4	1	3	2
	f1score	4	2	1	3
lichenoid	accuracy	4	3	1	2
	sensitivity	1	4	3	2
	specificity	4	2	1	3
	precision	3	4	1	2
	f1score	2	4	1	3
low-risk	accuracy	3	3	2	1
	sensitivity	3	2	4	1
	specificity	3	4	1	2
	precision	4	3	2	1
	f1score	4	3	2	1
high-risk	accuracy	4	2	1	3
	sensitivity	4	3	1	2
	specificity	1	3	2	4
	precision	4	2	1	3
	f1score	4	3	1	2
Aggregate rank (sum)		66	53	37	42
Final rank		4	3	1	2

6.5.2.3. Approach 3: Spatial distribution

All mean and standard deviations for all distances between the nuclei in each image were calculated and assessed using summary statistics (Figure 6.14.).

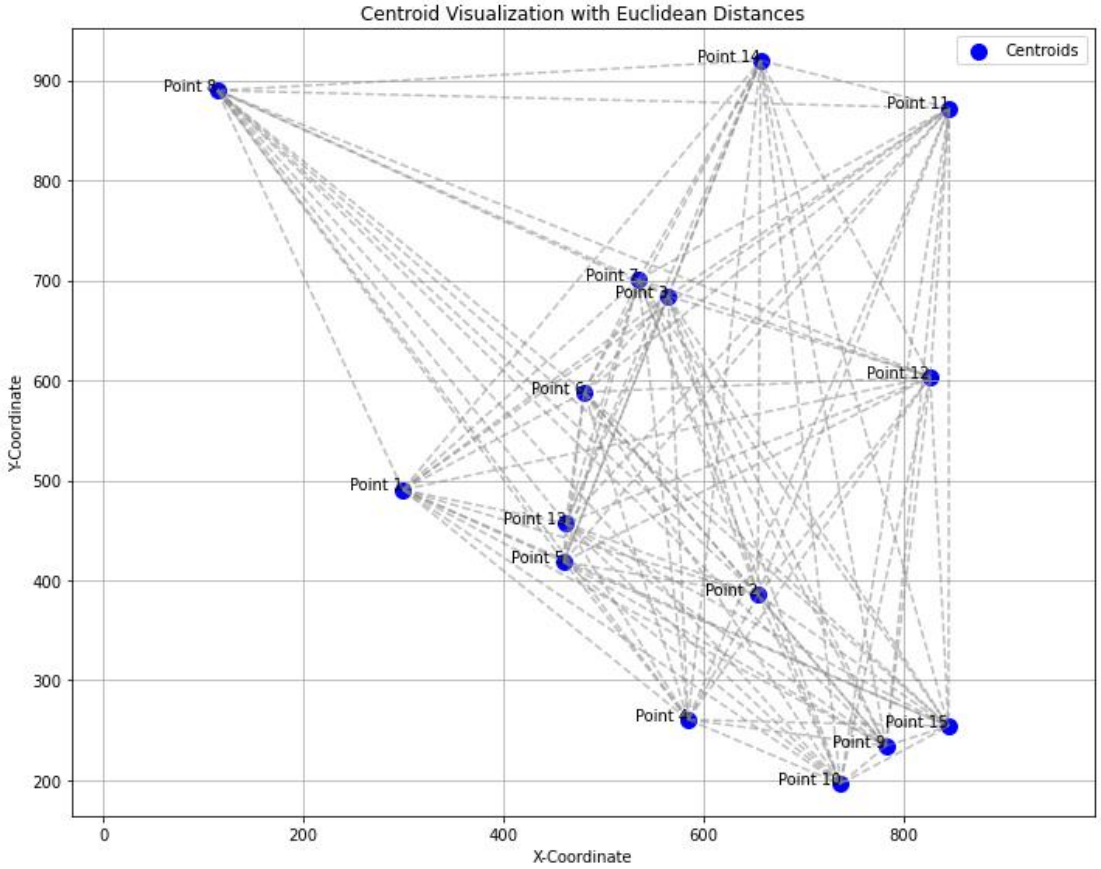


Figure 6.14. Graphical representation of nucleus distance measurements for an example fluorescein image

The no dysplasia images showed the highest mean and standard deviation values on average compared to all other diagnostic categories. The low-risk images showed the lowest mean distance. High-risk images while having the second largest distances on average showed the lowest variability. This indicates the no dysplasia images tended to have more spaced-out nuclei with a larger variability in their inter-nuclei distance. Low-risk images tended to have nuclei closer to each other. High-risk images while having relatively large distances between nuclei, had comparatively lesser variance in their distance measurements (Table 6.34.).

The one-way ANOVA of these measurements on fluorescein showed no significant differences in the mean and standard deviation of nuclei distances between the groups (Table 6.34.).

Table 6.34: Summary of means and standard deviation of distances between all fluorescein nuclei per image

		Normal	**Lichenoid**	**Low-risk**	**High-risk**	**P value for ANOVA**
mean distance	mean	485.54	477.01	477.15	486.85	0.781
	median	495.81	477.39	476.52	491.62	
	s.d.	97.23	81.42	119.44	100.12	
s.d. of distance	mean	227.27	223.18	229.35	222.52	0.728
	median	238.48	231.72	229.88	228.17	
	s.d.	52.41	52.52	65.47	52.08	

In terms of ML models developed on these two distance measurements the Random Forest (RF) model was the best performer in this approach, achieving the highest overall accuracy and F1-score. For the no dysplasia category, the RF model achieved an accuracy of 66%, sensitivity of 0.304, and F1-score of 0.292, indicating lower performance compared to the other approaches. The lichenoid category showed an accuracy of 57%, sensitivity of 0.389, and F1-score of 0.394 (Tables 6.35. & 6.36.).

Table 6.35. Approach 3 test results for the 4 ML models with respect to each diagnostic category (1 vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	0.75	0.54	0.66	0.58
	sensitivity	0	0.35	0.30	0.39
	specificity	0.97	0.60	0.77	0.64
	precision	0	0.21	0.28	0.24
	f1score	0	0.26	0.29	0.3
lichenoid	accuracy	0.43	0.53	0.57	0.52
	sensitivity	0.92	0.53	0.39	0.53
	specificity	0.16	0.53	0.67	0.52
	precision	0.38	0.39	0.4	0.38
	f1score	0.54	0.45	0.39	0.44
low-risk	accuracy	0.72	0.67	0.65	0.69
	sensitivity	0.16	0.08	0.28	0.08
	specificity	0.91	0.87	0.77	0.89
	precision	0.36	0.17	0.29	0.2
	f1score	0.22	0.11	0.29	0.11
high-risk	accuracy	0.84	0.84	0.72	0.83
	sensitivity	0	0	0.13	0.06
	specificity	1	1	0.83	0.98
	precision	0	0	0.13	0.33
	f1score	0	0	0.13	0.11

The low-risk category had an accuracy of 65%, sensitivity of 0.280, and F1-score of 0.286, suggesting the models struggled to classify low-risk nuclei in this approach. The high-risk category remained challenging, with the RF model achieving an accuracy of 72%, sensitivity of 0.125, and F1-score of 0.125 (Tables 6.35. & 6.36.). Overall, the

models performed poorly with classifying the approach 3 fluorescein nuclei distance measurements.

Table 6.36. Approach 3 ranks for the 4 ML models with respect to each diagnostic category (1 vs all) in the fluorescein test images

Diagnostic category	Metric	Logistic regression	SVM	Random Forest	XGBoost
no dysplasia	accuracy	1	4	2	3
	sensitivity	4	2	3	1
	specificity	1	4	2	3
	precision	4	3	1	2
	f1score	4	3	2	1
lichenoid	accuracy	4	2	1	3
	sensitivity	1	2	4	2
	specificity	4	2	1	3
	precision	4	2	1	3
	f1score	1	2	4	3
low-risk	accuracy	1	3	4	2
	sensitivity	2	3	1	3
	specificity	1	3	4	2
	precision	1	4	2	3
	f1score	2	4	1	3
high-risk	accuracy	1	1	4	3
	sensitivity	3	3	1	2
	specificity	1	1	4	3
	precision	3	3	2	1
	f1score	3	3	1	2
Aggregate rank (sum)		46	54	45	48
Final rank		2	4	1	3

6.5.2.4. Comparison of best fluorescein feature extraction ML model results

For fluorescein approach 1 of directly applying the ML models to the mean measurement valued provided the best results. The high-risk category again produced the greatest challenge in classification. The approach 1 model (random forest) showed a higher F1 score for not dysplasia (0.54), lichenoid (0.55), and high-risk (0.67), while having a marginally lower F1 score for low-risk (0.4) compared to the approach 2 model (random forest) (0.46) (Tables 6.37. & 6.38.). The high-risk category remained the most challenging for all approaches and models, indicating that the nuclei-level features may not be sufficient to accurately classify this category.
The random forest models were overall the best performing models across both contrast agent measurement datasets (Table 6.38.). Despite there being statistically significant differences between the nucleus measurements the ML models performed only moderately at being able to use those measurements to diagnose lichenoid lesions, OED, and OSCC.

The fluorescein approach 1 random forest model ranked the best across all other models (Table 6.37.). This model identified about 42.9% of low-grade OED lesions and about 55.6% high-grade OED and OSCC lesions (Table 6.38.).

Table 6.37: Test results of the best ranked ML models based on fluorescein segmentation data for all 3 analysis approaches

Diagnostic category	Metric	Best approach 1 model (RF)	Best approach 2 model (RF)	Best approach 3 model (RF)
no dysplasia	accuracy	0.70	0.65	0.66
	sensitivity	0.67	0.56	0.30
	specificity	0.71	0.68	0.77
	precision	0.46	0.40	0.28
	f1 score	0.54	0.47	0.29
lichenoid	accuracy	0.63	0.73	0.57
	sensitivity	0.50	0.50	0.39
	specificity	0.74	0.83	0.67
	precision	0.62	0.55	0.40
	f1 score	0.55	0.52	0.39
low-risk	accuracy	0.84	0.75	0.65
	sensitivity	0.43	0.42	0.28
	specificity	0.90	0.86	0.77
	precision	0.38	0.50	0.29
	f1 score	0.40	0.46	0.29
high-risk	accuracy	0.91	0.89	0.72
	sensitivity	0.56	0.57	0.13
	specificity	0.98	0.96	0.83
	precision	0.83	0.75	0.13
	f1score	0.67	0.65	0.13

Table 6.38. Ranking of best models across all approaches

Diagnostic category	Metric	Best approach 1 model (RF)	Best approach 2 model (RF)	Best approach 3 model (RF)
no dysplasia	accuracy	1	4	3
	sensitivity	3	5	6
	specificity	2	3	1
	precision	3	4	6
	f1score	4	5	6
lichenoid	accuracy	5	3	6
	sensitivity	1	1	5
	specificity	5	3	6
	precision	1	3	5
	f1score	1	2	5
low-risk	accuracy	1	4	6
	sensitivity	3	4	5
	specificity	2	3	6
	precision	4	3	5
	f1score	4	3	5
high-risk	accuracy	4	5	6
	sensitivity	2	1	4
	specificity	3	5	6
	precision	2	3	6
	f1score	1	2	4
Aggregate rank		52	66	102
Final rank		1	2	3

6.6. Top performing feature extraction machine learning model

The top performing model for the acriflavine dataset was the random forest (RF) model using the clustering and feature selection approach (approach 2) (Table 6.32.). For the fluorescein dataset the RF model using the mean measurements approach (approach 1) performed best.

Table 6.39. Best model performance in identifying OED and OSCC across all approaches using quantitative nuclei measurements

	Contrast agent	**Cases correctly identified** (sensitivity)	**Other categories wrongly predicted as this category** (1-specificity)	**Predictions of this category that were incorrect** (1 - precision)	**Cases misclassified** (1-sensitivity)
Low-grade OED	Acriflavine	58.80%	15.90%	42.90%	41.20%
	Fluorescein	42.90%	10%	62.50%	57.10%
High-grade OED & OSCC	Acriflavine	5.90%	1.20%	75%	94.10%
	Fluorescein	55.60%	2.10%	16.70%	44.40%

For low-grade OED, the acriflavine model identified more cases compared to its fluorescein counterpart (58.8% vs 42.9%). However, both models recommended several incorrect non-urgent biopsies and misclassified about half of the low-grade OED cases (Table 6.39.). For high-grade OED & OSCC, the fluorescein model identified about half of the cases (55.6%) whereas the acriflavine nearly missed all the cases (5.9%). Thus, the fluorescein RF model was considered to be the best performing feature extraction machine learning model.

Table 6.40. Test confusion matrix for best performing model (Fluorescein approach 1 - random forest) for diagnostic classification of quantitative nucleus features

		Predicted			
		No dysplasia	**Lichenoid**	**Low-risk**	**High-risk**
Actual	**No dysplasia**	10	5	0	0
	Lichenoid	8	13	5	0
	Low-risk	2	1	3	1
	High-risk	2	2	0	5

The test results for the best segmentation model per class are:

No dysplasia:

For the no dysplasia class, sensitivity was 0.67 and specificity 0.71, with modest precision (0.46) and an F1 score of 0.54. Misclassification frequently occurred with lichenoid lesions, where several true non-dysplastic cases were predicted as lichenoid.

Lichenoid:

The lichenoid group demonstrated balanced but moderate performance, with sensitivity of 0.50, specificity of 0.74, precision of 0.62, and an F1 score of 0.55. A substantial proportion of lichenoid cases were correctly identified, although overlap remained with both non-dysplastic and low-risk dysplasia categories.

Table 6.41. Test performance for best performing model (Fluorescein approach 1 - random forest) for diagnostic classification of quantitative nucleus features

Class	Sensitivity	Specificity	Precision	F1 score
No dysplasia	0.67	0.71	0.46	0.54
Lichenoid	0.50	0.74	0.62	0.55
Low-risk	0.43	0.90	0.38	0.40
High-risk	0.56	0.98	0.83	0.67

Low-risk:

In the low-risk dysplasia category, sensitivity was lower at 0.43 despite relatively high specificity (0.90). Precision was 0.38 and the F1 score 0.40, indicating the model

struggled to consistently identify low-risk lesions, often misclassifying them as non-dysplastic or lichenoid.

High-risk:

The high-risk dysplasia group yielded comparatively stronger results, with sensitivity 0.56, specificity 0.98, precision 0.83, and the highest F1 score (0.67) among all categories. This reflects good discriminative ability for high-risk lesions, though some cases were still misclassified as either no dysplasia or lichenoid.

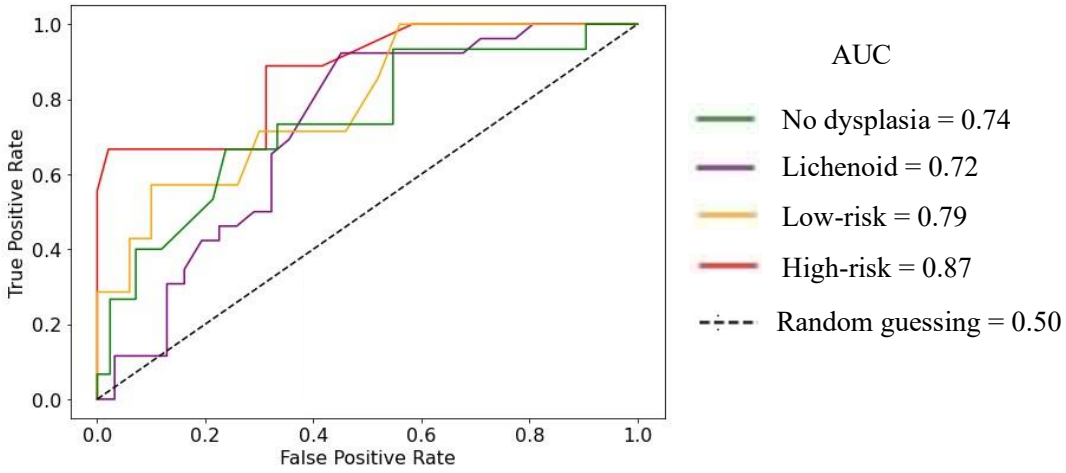


Figure 6.15. The AUROC results of test predictions for each diagnostic class by the best feature extraction ML model (fluorescein random forest approach 1 model)

This best performing feature extraction model had moderate AUC scores across all diagnostic categories with the highest score for high-grade OED and OSCC (0.87) followed by low-grade OED (0.79), no dysplasia (0.74) and lichenoid (0.72) (Figure 6.15.).

6.7. Discussion

As a result of the previous analyses (Chapter 5) assessing human annotated cellular and nuclear features with histopathological diagnoses of oral epithelial dysplasia and oral squamous cell carcinoma that demonstrated very poor sensitivity and specificity, we aimed to develop and validate a quantitative analysis method using machine learning. Quantitative feature extraction has been reported in the literature using a several different approaches (Ramani et al., 2023).

This segmentation technology used in our quantitative nucleus analysis study extracted the size, shape, brightness and spatial distribution of nuclei in our CLE images was paired with classification ML models to provide a predictive diagnosis of lichenoid, OED and OSCC. The first experiment in the present study compared the performance of two U-Net inspired deep learning segmentation algorithms, Cellpose 2D and StarDist 2D, for separating nuclei from the background within fluorescent in vivo confocal microscopy images (Stringer et al., 2021; Weigert et al., 2020).

The process of annotating the nuclei in the micrographs to prepare this data and the ground truth for training and testing the segmentation models was undertaken using Cellpose 2D and ImageJ software. Cellpose 2D allowed for the development of a bespoke human-in-the-loop annotation with the assistance of their default Cyto3 segmentation model that vastly reduced the time required to demarcate hundreds of nuclei in several hundred images. Human annotation of medical images is a labour-intensive, time-consuming, and expensive process (Zhang & Qie, 2023). This is particularly true for the current study, which required meticulous screening of microscopy images to identify subtle changes in nuclear structures of the oral epithelium. Further, expert annotations in medical imaging can often differ, resulting in potentially inconsistent labels (Yang et al., 2023). Disagreements regarding the presence or extent of findings, such as the nuclear features in the confocal microscopy images used in this study, may introduce noise into the training data, thereby affecting the performance of machine learning models (Koçak et al., 2025). Thus, the assistive annotation protocol using Cellpose 2D developed in this study can standardise and streamline the process of annotating and segmenting hundreds of thousands of objects on microscopy images.

Upon training and testing the StarDist 2D model performed significantly better ($p>0.015$) than Cellpose 2D at accurately marking the nuclei. Subsequently this development process was upscaled using the better performing StarDist 2D for optimising, training, and testing. The assessment of the segmentation performance of this model did not involve using the Intersection over Union metric (IoU) in its raw form due to specific limitations. For instance, when there is no overlap between predicted and ground truth boxes, IoU always returns 0, regardless of how close the

boxes are. While other metrics provide a more holistic evaluation, such as mean true score (MTS) that measures how well the model's segmentation aligns with the ground truth and mean matched score (MMS) that reflects how well the model's predicted segments align with its ground truth counterpart when there is an overlap. The developed StarDist 2D model segmented nuclei in the CLE micrographs with a high median MMS of 0.85 but a poor median MTS of 0.31. The high MMS confirms that the majority of the predicted nuclei present in the ground truth were accurately segmented. However, the poor MTS indicates that several ground truth nuclei had been missed. Since MTS is a scoring system that assigns a value of 0 for every missed prediction it brings down the overall score for any missed nuclei.

In our experiment that involved measuring nuclei, the correct segmentation of predicted nuclei was considered more important than under-segmentation as a sufficient number of measurements could be obtained from correctly identified nuclei for diagnostic predictions. The development of this trained nucleus segmentation model allowed for the segmentation of nuclei in the entire acriflavine and fluorescein datasets with high levels of confidence for the subsequent extraction of nuclei measurements. The measurement metrics were area, mean pixel intensity, standard deviation of pixel intensity, integrated density, circularity, and aspect ratio. While nucleus area described the mean number of pixels contained in a single nucleus, the mean pixel intensity and standard deviation illustrated the average pixel brightness values on a scale of 0-255 (8-bit scale) with 0 being completely black and 255 completely white. Integrated density combined area and mean pixel grayness to provide the average sum of all pixel brightness values contained in a single nucleus. Circularity and aspect ratio are shape descriptors that depict how round and oval the identified nuclei were.

The trained StarDist 2D model identified 91,550 nuclei across 1322 acriflavine images (21 images were excluded for having zero identifiable nuclei). Quantitative analysis of nuclear features revealed distinct characteristics across different diagnostic categories, with statistical significance observed in various comparisons. The Tukey's pairwise comparison post-hoc test provided valuable insights into these differences. Nuclear size, measured by area, showed limited discriminatory power, with significant differences only between lichenoid and low-risk nuclei ($p=0.029$). However, brightness-related features proved more informative. The normal nuclei exhibited the lowest overall fluorescence intensity, while high-risk nuclei were the brightest, showing the highest median mean grayness and integrated density. Mean pixel grayness effectively differentiated high-risk and low-risk nuclei from normal and lichenoid nuclei ($p<0.001$), indicating its potential as a diagnostic marker. Variance in brightness of nuclei, represented by the standard deviation of pixel grayness metric, emerged as a particularly informative measure. Significant differences were observed between normal, lichenoid, and low-risk nuclei, as well as between high-risk and low-risk nuclei ($p<0.001$). Notably, low-risk nuclei showed the greatest variability in brightness.

These findings suggest that fluorescence variance could be a key feature in distinguishing between diagnostic categories. Shape-based measurements also showed some diagnostic value. Nucleus circularity values significantly differed between normal and lichenoid nuclei ($p<0.001$), normal and high-risk nuclei ($p=0.029$), and lichenoid and low-risk nuclei ($p=0.016$). Additionally, normal nuclei had significantly higher aspect ratio values and thus tended to be more oval compared to lichenoid ($p<0.001$) and low-risk nuclei ($p=0.013$). Integrated density measurements further supported the use of brightness-related features in differentiation. Normal nuclei had significantly lower integrated density compared to low-risk ($p=0.038$) and high-risk ($p=0.019$) nuclei, with the high-risk nuclei also being significantly brighter than their lichenoid ($p=0.049$) counterparts.

The trained StarDist 2D model identified 10,722 nuclei across 640 fluorescein images, while 337 images were excluded for having zero identifiable nuclei. It is interesting to note that more than half of the fluorescein dataset did not having a single identifiable nucleus. Analysis of nuclear features in fluorescein-stained samples revealed distinct characteristics across different diagnostic categories, with statistical significance observed in most measurement metrics as determined by one-way ANOVA. Nuclear size, measured by area, showed significant variations. Low-risk nuclei were generally smaller, having the lowest median area ($p<0.001$). In contrast, high-risk nuclei were notably larger, with the highest median area ($p<0.001$). The Tukey's post-hoc test confirmed that high-risk nuclei area values were significantly different from all other categories, while lichenoid were significantly larger than low-risk nuclei ($p=0.043$).

The fluorescein nuclei brightness-related features provided valuable insights with different implications to the acriflavine set. Normal nuclei exhibited the highest mean pixel brightness, while lichenoid nuclei had the lowest values ($p<0.001$). Interestingly, lichenoid nuclei also showed the lowest pixel brightness standard deviation and integrated density, indicating the least fluorescence brightness overall ($p<0.001$). Fluorescence variance analysis, represented by the standard deviation of pixel brightness, showed significant differences between all pairs of diagnostic categories except for normal vs lichenoid. Notably, high-risk nuclei had the highest median brightness standard deviation, suggesting greater variability in brightness within these nuclei ($p<0.001$). Integrated density measurements further supported the distinctiveness of high-risk nuclei, as their values were significantly higher compared to normal ($p=0.013$), lichenoid ($p<0.001$), low-risk ($p=0.001$) indicating they were overall the brightest collection of nuclei. This, combined with their larger area and higher grayness standard deviation, indicates that high-risk nuclei were on average larger and brighter, whilst also displaying the most variable brightness. Shape-based measurements (circularity and aspect ratio) were relatively consistent across all diagnostic categories, suggesting that all segmented nuclei were fairly circular. Interestingly, circularity was the only metric that did not show statistically significant differences in the one-way

ANOVA, suggesting its limited utility as a discriminatory feature in fluorescein-stained confocal micrographs.

High-risk nuclei consistently exhibited high brightness (integrated density) across both contrast agent datasets, potentially serving as a reliable marker for high-grade OED and OSCC. This increase in pixel brightness due to increased concentration of contrast agent molecules in high-grade OED and OSCC nuclei is consistent with the nuclear hyperchromatism that is characteristic of dysplastic nuclei in histopathology and immunohistochemistry as noted in the WHO grading (Odell et al., 2021a). Similarly, lichenoid nuclei consistently showed low brightness and low variability in both contrast agent datasets, suggesting another stable characteristic for identification. However, the features of normal and low-risk nuclei were less consistent between contrast agents, indicating a need for caution when using these categories for classification.

Interestingly, nuclear size (area) emerged as a more distinguishing feature in the fluorescein dataset, particularly for high-risk nuclei. Shape measurements were generally less informative in both contrast agent datasets. The contradictory results for normal nuclei between the two contrast agent datasets, particularly in terms of brightness, highlight the potential impact of the different stains on the outcomes. It is important to note that it is difficult to draw comparisons gained from both contrast agent datasets regarding the nucleus measurement due to the large disparity in the number of nuclei identified and measured between them with the acriflavine dataset having close to 9 times more nuclei measured than its fluorescein counterpart. Background fluorescence intensity outside the segmented nuclei was not assessed in this study, as all non-nuclear regions were excluded during image preprocessing. The smaller number of high-risk nuclei available for analysis may also have influenced the apparent differences in staining intensity observed in this group.

This comprehensive analysis of the nuclear measurements in stained confocal micrographs suggested that a combination of nuclear features, including brightness, texture, and to some extent size and shape, could serve as valuable marker for differentiating between diagnostic categories. Inputting these measurements as features for the different diagnostic categories into ML models was therefore the next step in creating the diagnostic triage pipeline. The quantitative analysis of this measurement data with ML models to provide diagnostic predictions involved 3 different approaches. Directly inputting nuclei measurements was not possible as images had a variable number of identified nuclei and inputs to ML models need to feature vectors of a fixed standard size (fixed number of values).

Thus, the first approach (approach 1) summarised the nuclear measurements for each image by including the mean values for all measurements across all acriflavine and fluorescein datasets as input features for the 4 ML models (LR, SVM, RF, and XGBoost) with the output being the predicted diagnostic category. For the acriflavine dataset using nuclei measurement means, the SVM model performed best. However, its

performance varied across nuclei belonging to different diagnostic categories. The model performed best in identifying low-risk cases, with the highest accuracy (75.2%) and F1-score (0.554) among all categories. This suggests that low-risk nuclei have more distinct characteristics in their measurement means.

The model showed moderate performance for normal and lichenoid categories, with accuracies of 63.9% and 69.8% respectively. This suggests that these categories have some distinguishing features, but there's still room for improvement. The SVM model struggled most with high-risk cases, showing low sensitivity (0.20) and precision (0.33) potentially due to the high-risk nuclei measurement data sharing similar characteristics with other categories, making them harder to distinguish. This is in direct contradiction to the statistical results that indicated high-risk nuclei to be significantly brighter and larger. The fluorescein dataset's best approach 1 ML model was a RF type model which showed an overall superior classification performance compared acriflavine dataset SVM model with an accuracy across all categories ranging from 63.2% to 91.2%.

The fluorescein dataset model performed particularly well in identifying the critical high-risk category (accuracy = 91.2%, F1 score = 0.67) with a better balance between precision and recall. While its accuracy for identifying normal (70.2%), lichenoid (63.2%), and low-risk (84.2%) nuclei was high, the low F1 scores ranging from 0.4-0.55 indicated these models did not have high discriminatory power for these categories. Both of the Approach 1 contrast agent dataset ML models showed potential for the relevance of nuclei measurements as classification features. However, the overall moderate to poor performance indicates it may not be reliable for clinical use, especially for OED and OSCC detection.

Utilising the means of nuclei measurements in approach 1 therefore did not provide satisfactory results and so a subsequent method (approach 2) was developed to describe images using the actual nuclei measurements rather than the mean. A crucial requirement for all the ML models developed in this study was that they required the information representing each image to contain the same number of values. This was easily achieved in approach 1 since simply the mean value of each measurement metric represented each image by a vector of 6 numerical values. The goal with this second approach was to summarise the nuclei measurements in a different way such that the number of values representing the nuclei measurements for each image were the same.

To address this consideration k-means clustering algorithms were used to group nuclei with similar measurements as a way to summarise these measurements and describe each image with a fixed number of feature vector values. This feature vector consisted of cluster proportion values of nuclei in each image belonging to each cluster. K-means clustering is an unsupervised learning algorithm that groups datapoints that are close to each other in an n-dimensional space (6 dimensional in this case due to there being 6 measurement metrics). An important parameter of this clustering method is the number of clusters/groups needed to summarise the data appropriately. Elbow plots were used to

select the best values for number of clusters (k), which involve plotting the cohesiveness of clusters vs the number of groups used to cluster the data (Humaira & Rasyidah, 2020). Elbow plots for both acriflavine and fluorescein datasets indicated values of 3,4,5,6, and 7 were most appropriate in balancing within-cluster sum of squares (WCSS) (distance between datapoints in clusters) with the computational complexity of increasing the k value.

All these candidates for k-values were applied individually to create separate image cluster proportion feature vector datasets for the purpose of optimisation. On the feature selection front, Spearman's correlation matrices for both contrast agents showed nearly identical relationships between measurement features. Circularity and aspect ratio showed the strongest relationship (acriflavine = -0.92, fluorescein = -0.90). Moderately strong relationships were observed between integrated density vs area (acriflavine = 0.62, fluorescein = 0.69), integrated density vs mean pixel grayness (acriflavine = 0.73, fluorescein = 0.70), and mean pixel grayness vs standard deviation of grayness (acriflavine = 0.75, fluorescein = 0.56). All other correlations were poor, ranging from -0.26 to 0.41. The integrated density vs mean & area correlation values were expected because one method for calculating integrated density is a product of mean pixel brightness and area of the object. These observations assisted in identifying features that were too closely related and would potentially interfere in the impact of these features in clearly characterising the diagnostic categories. A sequential feature elimination of highly correlated features created datasets involving 5-features, 4-features, and 3-features including the original 6-feature set. Combining the k value datasets of 3,4,5,6 & 7 along with the 4 feature sets described above led to a total of 20 different combinations of datasets.

The same 4 ML model architectures as described previously were trained and tested on these datasets for both contrast agents. The RF model with the 5-feature set of mean pixel grayness, area, standard deviation of pixel grayness, integrated density, and aspect ratio along with clustering k=7 performed best compared to all other models on the acriflavine feature datasets. For the fluorescein datasets it was also the RF that outperformed all other ML architectures using the 4-feature set of mean pixel grayness, area, standard deviation of pixel grayness, and aspect ratio along with clustering k=6. Overall, the RF models in approach 2 showed relative consistency across both contrast agent datasets. For no dysplasia cases the acriflavine RF performed slightly better (accuracy=69.6%, sensitivity=0.745, F1-score = 0.66) compared to the fluorescein RF (accuracy=64.8%, sensitivity=0.56, F1-score=0.47). The acriflavine model (F1-score=0.58) marginally outperformed the fluorescein model (F1-score=0.45) for the low-risk class as well.

However, the fluorescein RF approach 2 model (F1-score=0.52) performed marginally better than its acriflavine counterpart (F1-score=0.49) for the lichenoid images. However, for the high-risk category the fluorescein approach 2 model (F1-score=0.65)

vastly outperformed the best acriflavine model (F1-score = 0.09). For acriflavine images the approach 2 (RF) model generally outperforms the previous SVM model, especially in accuracy and specificity. However, it struggled with sensitivity in the high-risk category (0.059 vs. 0.200 in the SVM model). On the other hand, the fluorescein approach 2 model was comparable to the approach 1 RF model while having slightly lower accuracy in the normal category (64.8% vs. 70.2%), better performance for the lichenoid class (73.0% vs. 63.2%) and similar performance in low-risk and high-risk categories.

Feature extraction approach 2 demonstrated a more consistent performance across both contrast agent datasets compared to previous models, suggesting its robustness in handling different data structures. However, a clear trade-off between sensitivity and specificity was observed, particularly in high-risk cases, where high accuracy is often accompanied by lower sensitivity, especially in the acriflavine dataset. This accuracy-sensitivity trade-off for high-risk nuclei is compounded by the dataset class imbalance where high-risk images constitute the smallest proportion of the entire dataset. The consistent performance of RF models across both datasets underscores the algorithm's suitability for the features present in nuclei. Despite an overall moderate performance, there remains significant room for improvement, particularly in balancing sensitivity and specificity for low- and high-risk cases. From a clinical perspective, while the high accuracy and specificity in OED and OSCC cases are promising, the low-moderate sensitivity suggests caution in relying on these models for critical diagnoses.

Pivoting away from the nuclear measurement metrics used in the first 2 approaches, the third quantitative analysis approach (approach 3) focused on measuring the distances between the nuclei, summarising these for the entire image in an attempt to quantify nucleus density. This is based on previous literature that highlights increased nuclear cell density in cancer cells (Uttam et al., 2015). This approach involved plotting the centroids of all identified nuclei, measuring the Euclidean distances between all nuclei with each other, and summarising all those measurements for each image with a mean and standard deviation. Upon measuring the nuclear distances in the acriflavine dataset, the high-risk images showed significantly smaller mean (p value all <0.001), and standard deviations (p value range 0.002-0.005) compared to the other groups. This agrees with previous literature and what is expected in high-grade OED and OSCC with abnormally increased mitosis, hyperplasia, and irregular epithelial cell architecture indicating that the cells themselves were tightly packed (Odell et al., 2021a). The nuclear distances in the fluorescein dataset showed no statistically significant differences between the diagnostic categories. This can be potentially attributed to the fluorescein dataset having 9 times fewer identified nuclei compared to the acriflavine dataset.

The same 4 ML model types as before were developed on this measurement dataset for both contrast agents independently. The best performing model for the acriflavine

dataset was the logistic regression (LR) model. The model shows high sensitivity (1.0) but low precision (0.39) for no dysplasia images, suggesting it over-predicts this class and likely misclassified many other types as no dysplasia. Despite the moderately high accuracy (78.7%) for the lichenoid class, the model failed to identify any true lichenoid cases (sensitivity = 0), indicating it's likely classifying all images as non-lichenoid. The model completely failed to identify any low-risk cases, despite moderate accuracy (67.2%), suggesting it's classifying all images as not low-risk. The model achieved very high accuracy (93.9%) but very low sensitivity (0.06) for high-risk cases, indicating that it is reasonable at identifying non-high-risk cases but poor at detecting actual high-risk ones. These results indicated that while high-grade OED and OSCC nuclei tended to be closer together and had significant variations in spatial distribution, the quantitative measurements of these features when paired with ML models did not provide results that are clinically applicable.

The random forest (RF) ML model performed best on the fluorescein dataset; however the performance was generally lower than the LR model on the acriflavine dataset. The RF model demonstrated moderate performance across all categories, with accuracies ranging from 57% to 72%. However, it struggled correctly identifying positive high-grade OED and OSCC cases in the fluorescein dataset. The poor performance in classifying fluorescein nuclei distance measurements in this approach indicates that these features may not be sufficiently highlighted by the contrast agent for accurate categorization across all dysplasia levels. These results highlight the challenges in developing robust machine learning models for dysplasia classification based solely on distance measurements and suggest the need for additional or more informative features to improve model performance, particularly for the clinically important low-risk and high-risk categories.

The best performing ML models across all approaches and both contrast agents were the fluorescein random forest model from approach 1. The direct input of means of pixel brightness and nuclear shape measurements provided the ML model with the best means to differentiate the diagnostic classes compared to the other approaches. The model showed lower sensitivity in the low-risk group, which highlighted the difficulty of distinguishing early dysplastic changes from non-dysplastic or lichenoid pathology, a limitation consistent with clinical practice where subtle morphological overlap often complicates diagnosis. For non-dysplastic and lichenoid cases, the moderate sensitivities suggested that the model captured some discriminative features, although frequent misclassifications between these categories indicated considerable overlap in their feature representation. By contrast, the high specificity observed for high-risk cases was encouraging as it reduced the risk of false positive alarms, yet the moderate sensitivity underscored the need to improve recall to minimise the chance of missing critical lesions.

This model had moderate to high AUC values for the diagnostic classes ranging from 0.72 to 0.87 (Figure 15). However, from a clinical perspective of predicting the need for biopsies for OED and OSCC the model fell short. It only correctly identified about half of the OED and OSCC cases. The models and the entire feature extraction approach described in the present study on the fluorescence in vivo confocal microscopy images was unfit for use in a clinical diagnostic triage setting for OED and OSCC.

Feature extraction brings the finer details of the imaging process into question due to the potential impact they can have on the performance of ML models. Difference in contrast agents are one such aspect that could influence quantitative feature extraction. Acriflavine binds to nucleic acids and fluorescein does not directly bind to nuclei but instead it is internalised within cells and is distributed within the cytoplasm and nuclei via passive diffusion (Piorecka et al., 2022; Robertson et al., 2013). This potentially had an impact on the differences in pixel brightness values between images of both contrast agents.

Lighting conditions and field of view (FOV) selection can influence the information available to feature extraction models. Variations in lighting can alter the appearance of features within an image, affecting the consistency and reliability of feature extraction. For instance, changes in illumination can introduce shadows or highlights that obscure or distort key features, leading to decreased model accuracy. To mitigate these effects, preprocessing techniques such as illumination normalization and data augmentation have been employed. However, the direct effect of external light illumination is on the brightness values of individual pixels which can affect metrics such as mean pixel brightness and integrated density. This effect cannot be mitigated in post-processing and needs to be addressed when the images are being captured.

Although this study addresses ‘feature-extraction’ it focuses primarily on nuclei. Other features such as cell borders, intra-cellular structures, and connective tissue elements were not extracted in the study due to clear appearance of nuclei, limitation of time and computing resources. However just like in histopathology analysis, nuclear appearance is just one diagnostic feature, and a host of other epithelial architectural changes are considered. The ability to visualise the epithelial architectural features in fluorescence microscopy heavily relies on the ability of the dye to penetrate and highlight the entire depth of the epithelium. This is challenging with topically applied contrast agents as was done in this study. Even in this study a few confocal micrographs showed an absence of nuclei which might allude to the limited efficacy of the topical fluorescence agent. These are situations where more targeted biomarker contrast agents such as Poly ADP-ribose Polymerase Inhibitor-Fluorescent (PARPi-Fl). It's a cutting-edge molecular imaging agent that combines the specificity of a DNA repair enzyme PARP1 with a fluorescent tag, typically a BODIPY-FL dye to specifically highlight DNA damage. This has been used to image advanced oral cancer cases (Demetrio de Souza Franca et al., 2021).

Despite the feature extraction method not working as intended, the overall findings from the nuclear quantitative measurement study identified patterns in the nuclei measurements across that might extend to cellular and architectural changes. This approach focused solely on nuclei but in order to consider all the other features present in these fluorescence confocal micrographs convolutional neural networks (CNN) were tested next.

7. DEEP LEARNING CLASSIFICATION DIAGNOSTIC TRIAGE MODELS

7.1. Introduction

Convolutional neural networks (CNN) are currently the best performing deep learning models for computer vision problems. Several image analysis tasks that had previously seemed impossible, such as facial recognition algorithms, guidance systems for autonomous vehicles, and produce identification at self-service supermarkets, are currently being solved using CNNs (Z. Li, Liu, Yang, Peng, & Zhou, 2021). CNNs are feedforward networks that can extract features from image data by convolving over the pixels. The design of a CNN is inspired by the biologic visual perception neurons (Z. Li et al., 2021).

Over the years, several CNN architectures have been proposed to improve image classification performance. One such model is the Inception_V3 CNN architecture model that performed exceptionally well having a top-5 error rate of 3.58% at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015 (Deng, Dong, Socher, Li, Li, et al., 2009; Szegedy et al., 2016). This model is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers (Szegedy et al., 2016).

Classifiers such as CNNs are induced from the datasets used to train them. Each classifier has an associated 'true' prediction accuracy that is unknown and cannot be calculated in real world problems (Rodriguez, Perez, & Lozano, 2009). The accuracy needs to be estimated from the available training data. There are several estimators of classification error, and the most popular one is k-fold cross validation (Stone, 1974). These estimators help configure the internal parameters of the CNN models which are learned from data, however deep learning models also have hyperparameters. These are external parameters that are not defined based on the dataset but are specified by a practitioner using a heuristic or trial and error until acceptable model accuracy is achieved (Alibrahim & Ludwig, 2021).

CNNs have several such parameters such as number of epochs (cycles of the model going through the entire training data) and learning rate. These hyperparameter values can be determined manually using previous examples or consulting experts. An alternate is to use tuning strategies that attempt to minimize the error of the algorithm using a search space of candidate configurations (Alibrahim & Ludwig, 2021). One such strategy is the grid search algorithm that is a systematic and exhaustive method to fully assess a search space attempting all possible combinations with the assumption that all parameters have the same probability of impacting the model performance (Marinov & Karapetyan, 2019).

Advancements in the field of deep learning and computer vision were accelerated due to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This involved a worldwide challenge for developing CNN models that could learn image classification

from over 14 million images across 20,000 categories in the ImageNet dataset (Deng, Dong, Socher, Li, Li, et al., 2009). This competition acted as a benchmark for performance of CNN models such as Inception_v3 developed by Szegedy et al. (2015) (Szegedy et al., 2016). Model architectures like Inception_v3 can be optimized for image classification tasks other than what they were created for using transfer learning. Transfer learning involves modifying the classification head layers of pre-trained neural networks in order to repurpose them for a specific image classification task while retaining their optimized weights and parameters from their original training session (Zhuang et al., 2020).

Aim: To develop, train and test convolutional neural networks on captured human oral in vivo confocal micrographs for the detection of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC).

7.2. Methods

This study describes a protocol for developing, training, and testing convolutional neural network (CNN) models for the detection of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) using fluorescence in vivo confocal microscopy images of the oral mucosa.

The research was conducted in two experiments, one using MATLAB's Deep Network Designer framework (MathWorks, U.S.A.) and another using the PyTorch machine learning framework in the python programming language, both employing transfer learning with the Inception_v3 convolutional neural network (CNN) architecture as the base template.

All images used for development, training and testing were categorised in the following diagnostic classes: (Figure 7.1., Table 3.2. from Chapter 3)

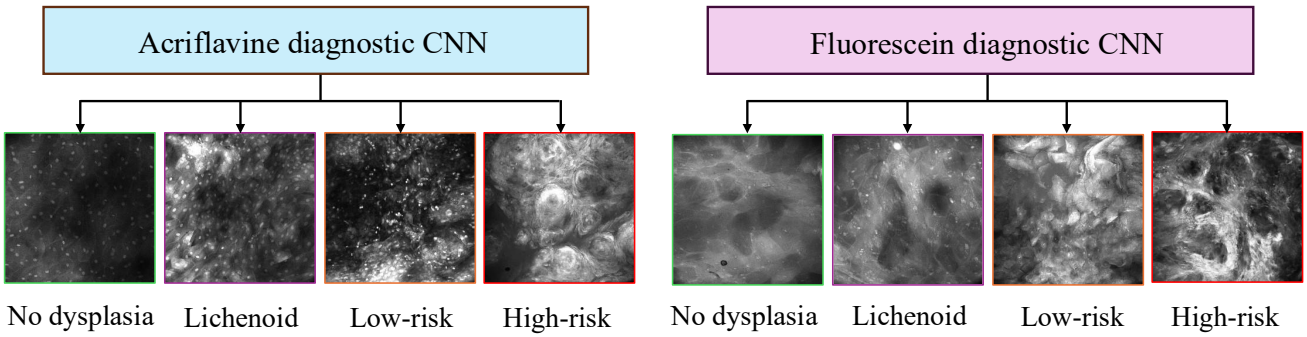


Figure 7.1. Examples of in vivo confocal micrographs across both contrast agents and all diagnostic categories used for CNN diagnostic triage

1. No dysplasia - included all lesions that do not show any signs of dysplasia such as amalgam tattoo, chronic inflammation, denture-associated hyperplasia, fibroepithelial polyp, focal papillomatosis, hyperplasia, hyperkeratosis, squamous papilloma, and verrucous xanthoma
2. Lichenoid – represented oral lichen planus and lichenoid inflammation
3. Low-risk – included low-grade oral epithelial dysplasia (OED) with atypia and verrucous hyperplasia
4. High-risk – contained all high-grade OED lesions and oral squamous cell carcinoma (OSCC).

The PyTorch QMR was applied to all 9168 available images, resulting in 1983 diagnostic quality images usable for analysis were used for training and testing the diagnostic triage CNN models. This diagnostic quality dataset of 1343 acriflavine and

640 fluorescein images were divided into 4 categories of ‘no dysplasia’, ‘lichenoid’, ‘low-risk’ and ‘high-risk’ across 80% training and 20% test images.

The dataset used for training and testing the CNN models in this study are depicted in Table 4.7. from Chapter 4. The training and test datasets for both contrast agents were divided with a ratio of 80:20.

CNNs were developed using MATLAB’s Deep Network Designer application and the PyTorch framework and Python 3 programming language. The CNN architecture was a modified Inception_V3 with transfer learning. The MATLAB CNN model was developed using the default set-up of 30 epochs and a learning rate of 0.001. Hyperparameter optimization for the PyTorch CNN model was performed using a grid search approach, exploring combinations of epochs (5,10,15) and learning rates (0.001, 0.01, 0.1). K-fold cross-validation (K=5) was employed across both contrast agent datasets, resulting in a total of 90 CNN models being trained and tested in PyTorch (Figure 7.2.). CNN performance was assessed using multiple evaluation metrics (Table 3.3. from Chapter 3):

1. Accuracy: The proportion of correctly classified instances out of the total instances.
2. Sensitivity (Recall): The proportion of actual positives correctly identified by the model.
3. Specificity: The proportion of actual negatives correctly identified by the model.
4. Precision: The proportion of predicted positives that are actually positive.
5. F1-score: The harmonic mean of precision and recall, balancing both metrics.
6. Received operator characteristic (ROC) Curve: A plot of the true positive rate (sensitivity) against the false positive rate across different thresholds.
7. AUROC (Area Under the ROC Curve): A single value summarizing the ROC curve, indicating the model's ability to distinguish between classes.

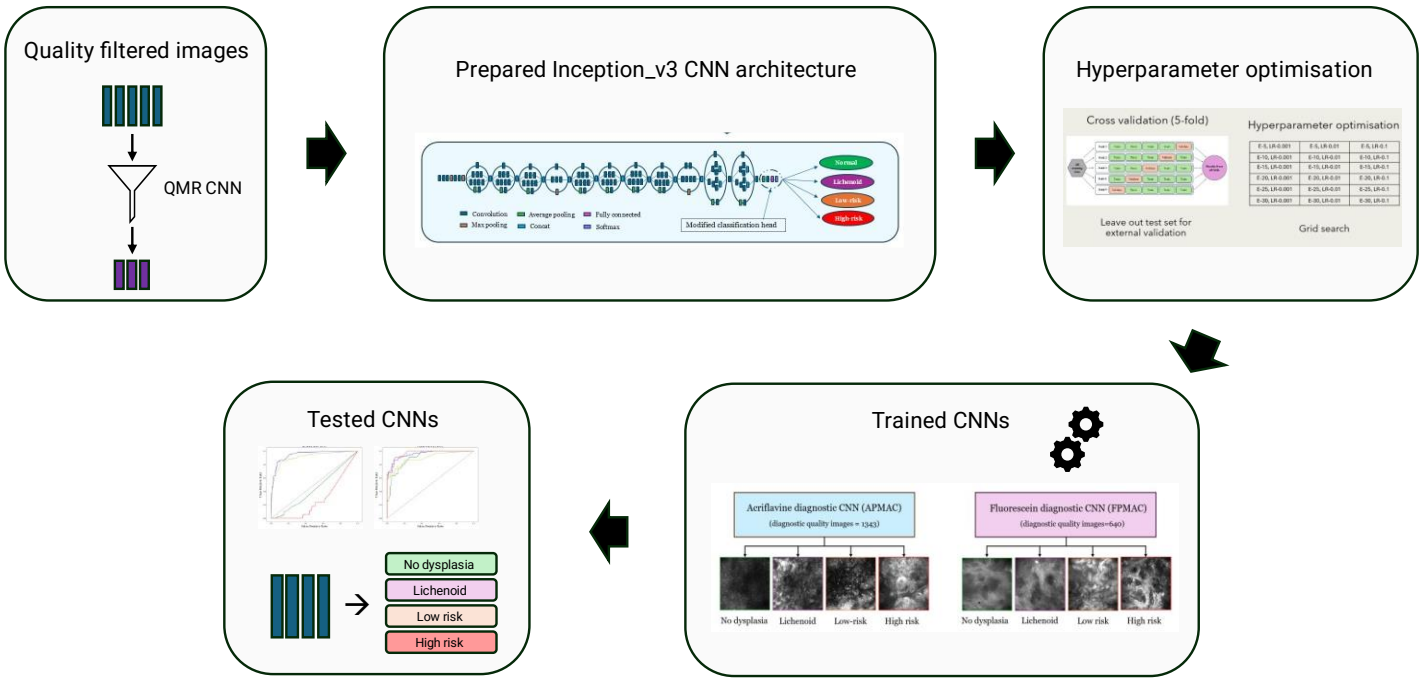


Figure 7.2. Diagnostic triage CNN development workflow

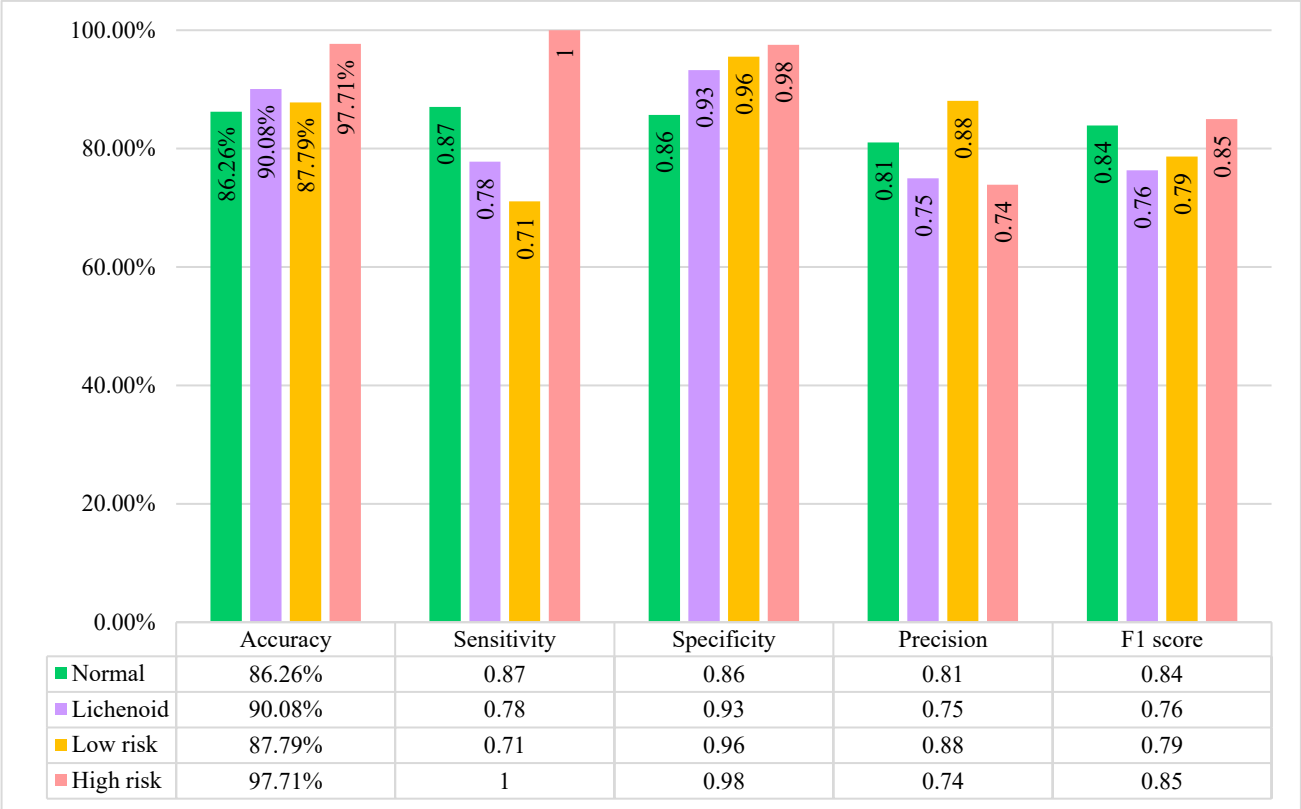
The PyTorch models were ranked based on aggregated performance across metrics and classes. These were compared with the MATLAB CNN. All Python and PyTorch code used in this study are available online at the provided GitHub repository (Appendix 2). Statistical analyses were performed using the Python sci-kit learn library.

7.3. Performance of diagnostic triage CNNs developed in MATLAB

7.3.1. Acriflavine MATLAB diagnostic CNN

The trained acriflavine diagnostic CNN model was tasked with predictions on the test images, and the model predictions were compared to the histopathology diagnosis. The predictions of 262 test images into their respective categories took the model 30.14 seconds (0.09 seconds per image).

The overall accuracy of the model was 80.91% with the highest per class accuracy for high-risk images (97.71%) followed by lichenoid (90.08%), low risk (87.79%) and no dysplasia (86.26%). Model sensitivity was moderate to high for the diagnostic classes with it being the least for low risk (0.71) and highest for high risk (1) (Figure 7.3.). Specificity was consistently high across all diagnostic categories ranging from 0.86-0.98 (Figure 7.3.). The model had moderate to high precision with high-risk classifications being the least precise (0.74) and low risk being the most precise (0.88). The F1 scores were moderate for lichenoid (0.76) and low-risk (0.79) images and high for no dysplasia (0.84) and high-risk (0.85) images. The F1 scores of the categories ranged from 0.76 to 0.85 (Figure 7.3.). The model identified all high-risk images correctly and only 6 other lesions were misclassified as high risk with 83.3% of them actually belonging to the low-risk category (Figure 7.3.).



	Accuracy	Sensitivity	Specificity	Precision	F1 score
Normal	86.26%	0.87	0.86	0.81	0.84
Lichenoid	90.08%	0.78	0.93	0.75	0.76
Low risk	87.79%	0.71	0.96	0.88	0.79
High risk	97.71%	1	0.98	0.74	0.85

Figure 7.3. MATLAB acriflavine diagnostic triage CNN test results for all diagnostic categories (1 vs all)

The acriflavine MATLAB diagnostic triage CNN model had high AUC values for lichenoid (0.96), and no dysplasia (0.95), followed by low-risk (0.94) and a perfect AUC for high-risk (1) on the test dataset (Figure 7.4.).

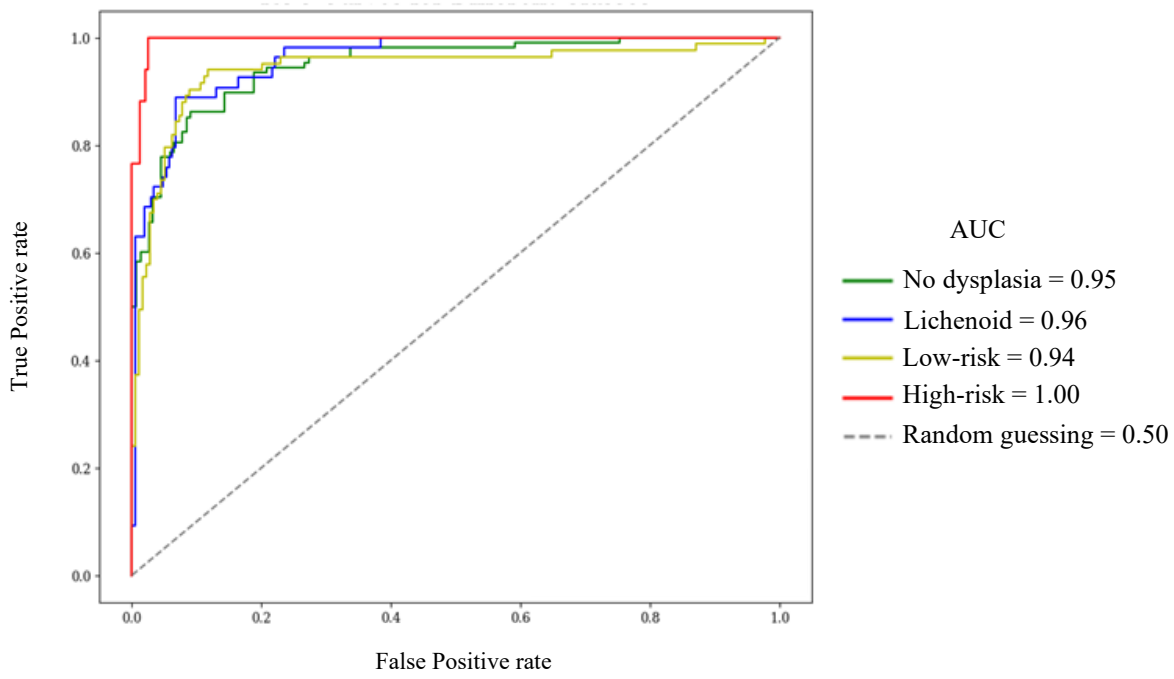


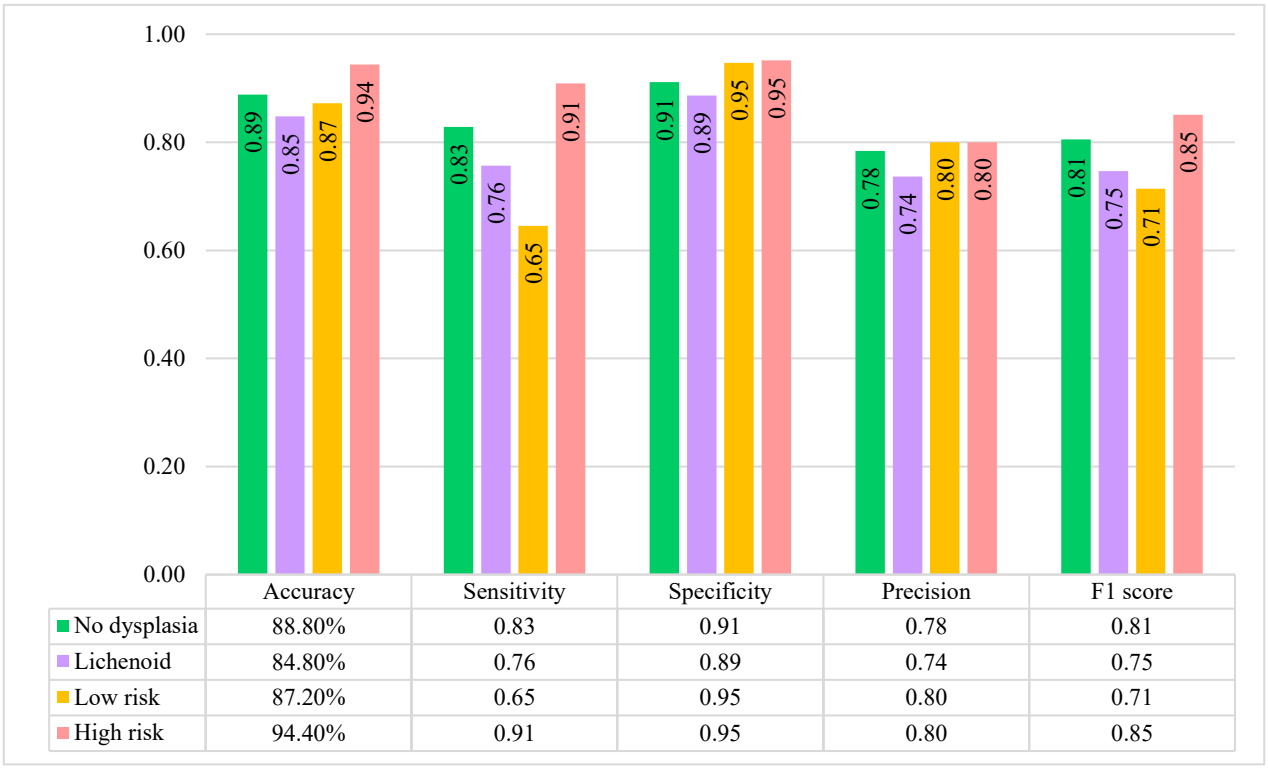
Figure 7.4. AUROC of the acriflavine MATLAB diagnostic triage CNN model

7.3.2. Fluorescein MATLAB diagnostic CNN

The trained fluorescein diagnostic CNN model was tasked with predictions on the test images, and the model predictions were compared to the histopathology diagnosis. This trained diagnostic triage CNN model was subsequently run on the test images, and the model predictions were compared to the histopathology diagnosis. The predictions of 125 test images into their respective categories took the model 21.43 seconds (0.14 seconds per image).

The overall accuracy of the model was 77.6% with the individual class accuracy being the highest for high-risk (94.4%) followed by no dysplasia (88.8%), low-risk (87.2%), and lichenoid (84.8%). Model sensitivity was moderate to high for the diagnostic classes with it being the least for low risk (0.65) and highest for high-risk (0.91) (Figure 7.5.). Specificity was consistently high across all diagnostic categories ranging from 0.89-0.95 (Figure 7.5.). The model had moderate to high precision with lichenoid classifications being the least precise (0.74) and low- & high-risk being the most precise (0.80). The F1 scores were moderate for no dysplasia (0.81), lichenoid (0.75) and low-risk (0.71) while being higher for high-risk (0.85) images (Figure 7.5.).

The fluorescein MATLAB diagnostic triage CNN model had high AUC values for high risk (0.98), followed by no dysplasia (0.94) with low-risk and lichenoid having the same score (0.91) (Figure 7.6.).



	Accuracy	Sensitivity	Specificity	Precision	F1 score
No dysplasia	88.80%	0.83	0.91	0.78	0.81
Lichenoid	84.80%	0.76	0.89	0.74	0.75
Low risk	87.20%	0.65	0.95	0.80	0.71
High risk	94.40%	0.91	0.95	0.80	0.85

Figure 7.6. MATLAB fluorescein diagnostic triage CNN test results for all 3 categories

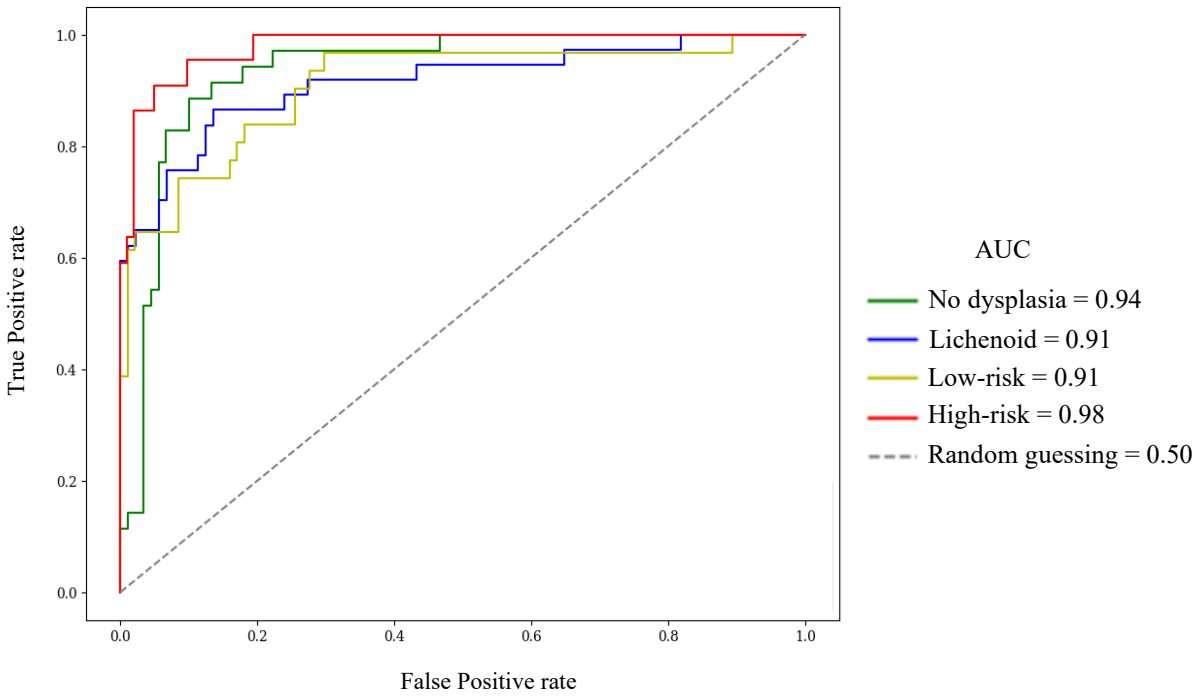


Figure 7.5. AUROC of the fluorescein MATLAB diagnostic triage CNN model

7.4. Performance of diagnostic triage CNNs developed in PyTorch

A total of 90 CNN models being trained and tested in PyTorch across all combinations of epochs, learning rates and cross validation folds for each of the 2 contrast agents. This resulted in a total of 180 diagnostic CNN models.

7.4.1. Acriflavine PyTorch diagnostic CNN

A total of 90 acriflavine PyTorch models were developed, trained and tested on 1081 and 262 images from the acriflavine diagnostic quality image set.

With a learning rate of 0.001, lichenoid image predictions consistently outperformed other classes, achieving an F1 score of 0.83 at 30 epochs, supported by balanced sensitivity (0.79) and high specificity (0.96). The models performed reasonably on low-risk images with an F1 score of 0.85, with high specificity (0.95) and recall (0.83). However, the models struggled with no dysplasia tissue images across all configurations, with F1 scores peaking at only 0.07 at 20 epochs due to poor sensitivity (0.04) and precision (0.19). High-risk predictions consistently showed the weakest performance, with F1 scores below 0.1 across all epochs, primarily due to low precision and sensitivity (Figure 7.7.).

When the learning rate increased to 0.01, the performance of the PyTorch acriflavine diagnostic triage CNN models fluctuated. Lichenoid maintained its relatively strong results, peaking at an F1 score of 0.78 at 25 epochs. Low-risk demonstrated moderate improvement, achieving an F1 score of 0.74 at 25 epochs, reflecting a better balance of precision and recall. However, model performance on no dysplasia images remained inadequate, and high-risk category results showed no significant improvement, with F1 scores consistently below 0.1. At the highest learning rate of 0.1, the performance deteriorated significantly across all classes. Although specificity remained high for no dysplasia and lichenoid classes, sensitivity and precision were too low to generate meaningful F1 scores. Low-risk and high-risk classes struggled similarly, with poor performance and F1 scores often near zero (Figure 7.7.).

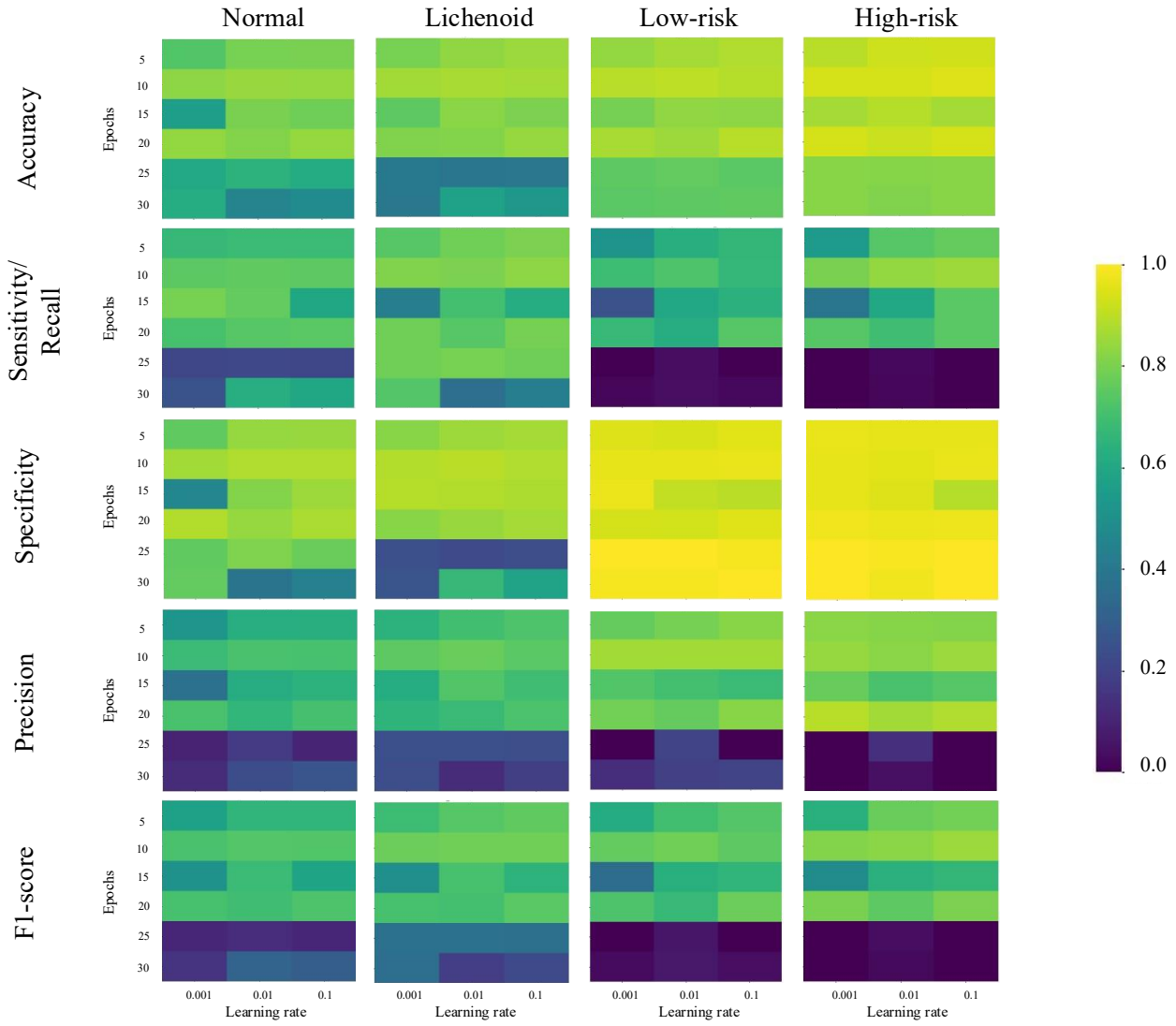
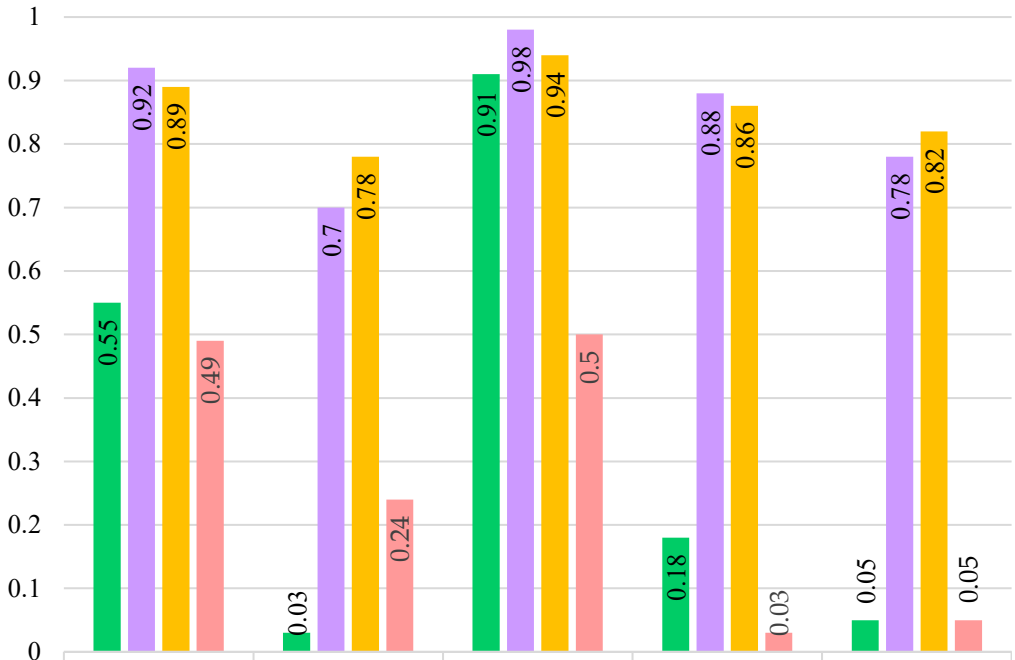


Figure 7.7. Heat maps depicting averaged performance metrics of PyTorch acriflavine diagnostic triage CNN for all 4 diagnostic classes across all hyperparameter combinations

The best ranked PyTorch acriflavine diagnostic triage CNN model was developed using 30 epochs with a learning rate of 0.001 (Figure 7.8.). The model showed strong accuracy for lichenoid (0.92) and low-risk images (0.89), but significantly lower values for no dysplasia (0.55) and high-risk (0.49). These discrepancies suggest inconsistent general performance across classes. Sensitivity was low across most classes, with no dysplasia (0.03) and high-risk (0.24) being particularly weak. Lichenoid (0.70) and low-risk (0.78) demonstrated better performance in detecting true positives. The model displayed strong specificity for lichenoid (0.98) and low-risk (0.94) and acceptable performance for no dysplasia (0.91). However, specificity for high-risk dropped to 0.50, indicating room for improvement in distinguishing true negatives in these images' vs the rest. Precision ranged from very low in high-risk (0.03) and no dysplasia (0.18) to high in lichenoid (0.88) and low-risk (0.86) images. The F1 score followed a similar trend,

being highest in lichenoid (0.78) and low-risk (0.82) but extremely low in high-risk (0.05) and no dysplasia (0.05) images.



	Accuracy	Sensitivity	Specificity	Precision	F1 score
No dysplasia	0.55	0.03	0.91	0.18	0.05
Lichenoid	0.92	0.7	0.98	0.88	0.78
Low-risk	0.89	0.78	0.94	0.86	0.82
High-risk	0.49	0.24	0.5	0.03	0.05

Figure 7.8. Test results of the best ranked PyTorch acriflavine diagnostic triage CNN model

The ROC curves for the acriflavine diagnostic triage CNN model further depict the disparity in the model's ability to detect lichenoid (AUC=0.94) and low-risk lesions (AUC=0.91) as opposed to no dysplasia (AUC=0.44) and high-risk (AUC=0.28) lesions (Figure 7.8.). The best ranked acriflavine diagnostic triage CNN model took 16.60 seconds to classify all 262 test images at the rate of 0.06 seconds per image.

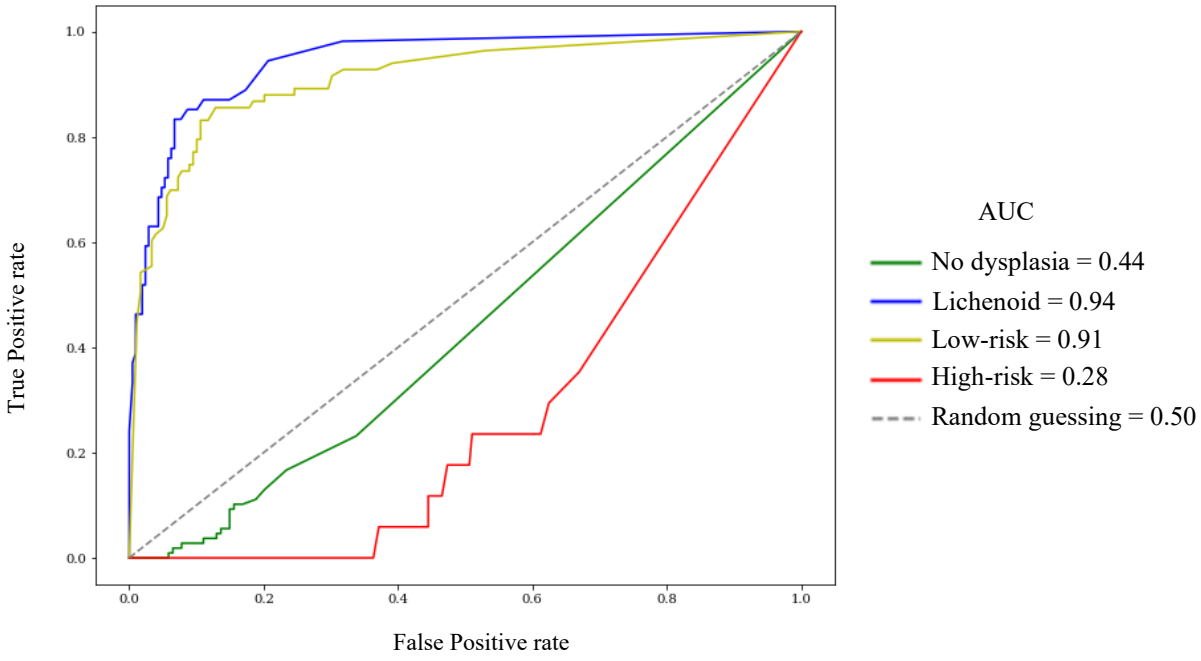


Figure 7.9. ROC curves for all diagnostic classes from the test results of the best ranked PyTorch acriflavine diagnostic triage CNN model

7.4.2. Fluorescein PyTorch diagnostic CNN

A total of 90 fluorescein PyTorch models were developed, trained and tested on 515 and 125 images from the fluorescein diagnostic quality image set.

The results of the PyTorch fluorescein diagnostic triage CNN were more consistent and demonstrated better overall performance compared to their acriflavine counterparts. At a learning rate of 0.001, all classes achieved stronger F1 scores, with high-risk class excelling at 30 epochs with an F1 score of 0.85, supported by high sensitivity (0.85) and specificity (0.97) (Figure 7.10.). The lichenoid class performed reliably well across all epochs, peaking at an F1 score of 0.83 at 30 epochs. The low-risk class also showed high performance, achieving an F1 score of 0.85 at 30 epochs, reflecting its strong precision and recall. While the results of the no dysplasia tissue class exhibited notable improvement compared to the PyTorch acriflavine diagnostic triage CNN models, its F1 score remained slightly lower than the other classes, peaking at 0.73 at 30 epochs (Figure 7.10.).

At a learning rate of 0.01, the results were still moderately high with the lichenoid and high-risk classes outperforming the others. High-risk prediction results peaked with an F1 score of 0.80 at 30 epochs, supported by high precision (0.85) and recall (0.75).

Low-risk results were moderate with F1 scores reaching 0.77 at 30 epochs. The lichenoid class showed consistent performance, achieving an F1 score of 0.78 at 30 epochs. The F1 score for the no dysplasia class images was lower (0.72) at 30 epochs. At the highest learning rate of 0.1, model performance dropped significantly, similar to the PyTorch acriflavine models. While specificity remained high for all classes, low sensitivity and precision severely impacted the F1 scores. Although low-risk and high-risk classes achieved marginally better results compared to the other classes, the overall performance was poor (Figure 7.10.).

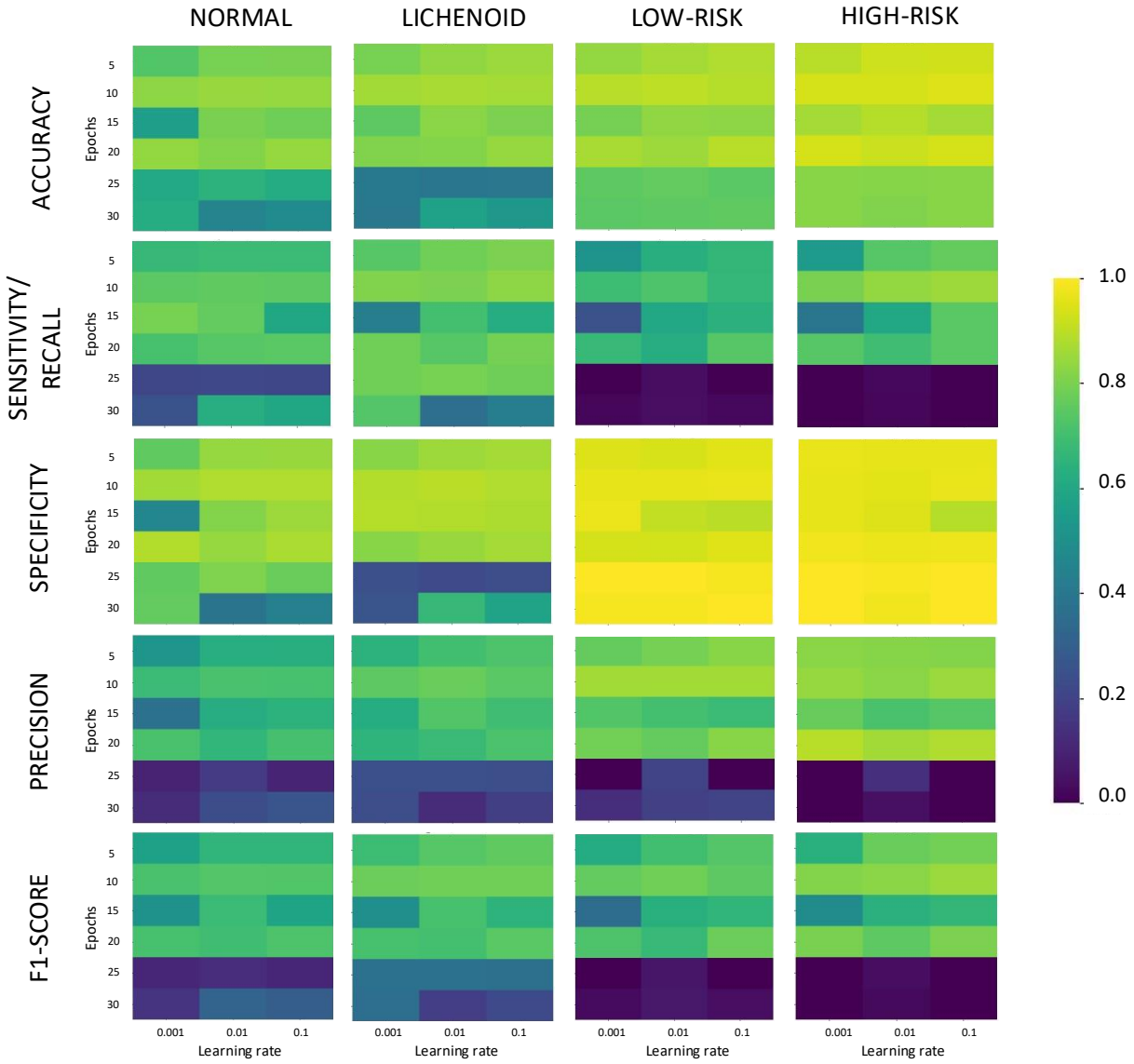
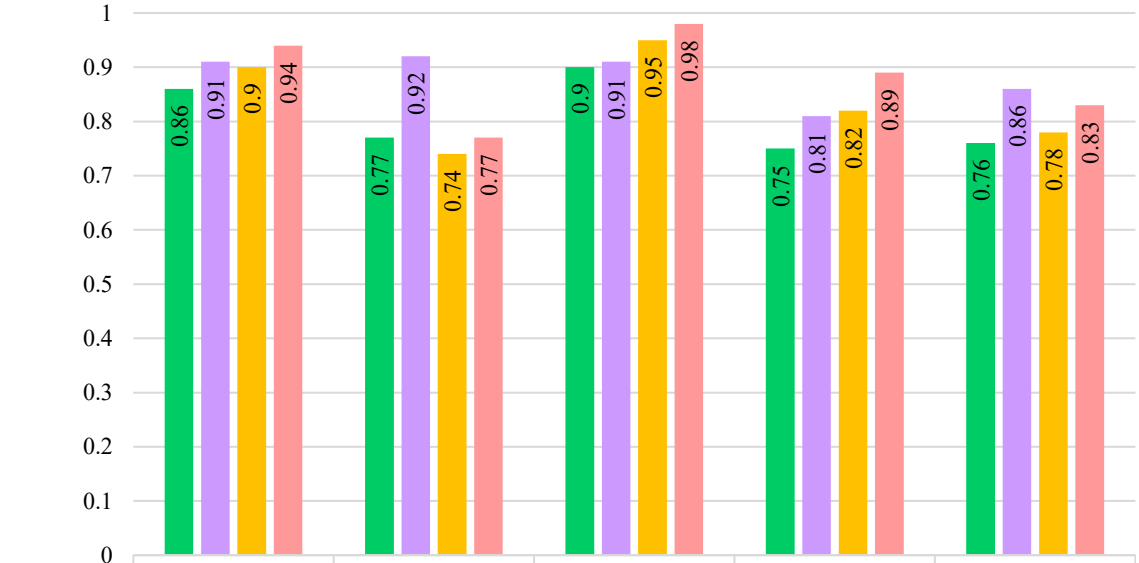


Figure 7.10. Heat maps depicting averaged performance metrics of PyTorch fluorescein diagnostic triage CNN for all 4 diagnostic classes across all hyperparameter combinations

The best ranked PyTorch fluorescein diagnostic triage CNN model was developed with 25 epochs and a learning rate of 0.001 (Figure 7.11.). This CNN achieved consistently high accuracy across all classes, ranging from 0.86 (no dysplasia) to 0.94 (high-risk). Sensitivity values were high overall with the lowest score being 0.74 (low-risk) and the

highest 0.92 (lichenoid). Specificity was consistently high, peaking at 0.98 for high-risk, suggesting exceptional ability in avoiding false positives with the lowest value of 0.90 for no dysplasia tissue. Precision ranged from 0.75 (no dysplasia) to 0.89 (high-risk), and F1 scores followed a similar trend, with lichenoid (0.86) and high-risk (0.83) achieving the highest values. This balance across metrics highlights strong, consistent performance (Figure 7.11.).



	Accuracy	Sensitivity	Specificity	Precision	F1 score
Normal	0.86	0.77	0.9	0.75	0.76
Lichenoid	0.91	0.92	0.91	0.81	0.86
Low-risk	0.9	0.74	0.95	0.82	0.78
High-risk	0.94	0.77	0.98	0.89	0.83

Figure 7.11. Test results of the best ranked PyTorch fluorescein diagnostic triage CNN model

The AUC for the ROC curves for no dysplasia (AUC=0.91), lichenoid (AUC=0.96), low-risk (AUC=0.90), and high-risk (AUC=0.96) shows the model effectiveness at identifying all the classes (Figure 7.12.). This fluorescein diagnostic triage CNN model took 5.59 seconds to classify all 125 test images at the rate of 0.04 seconds per image.

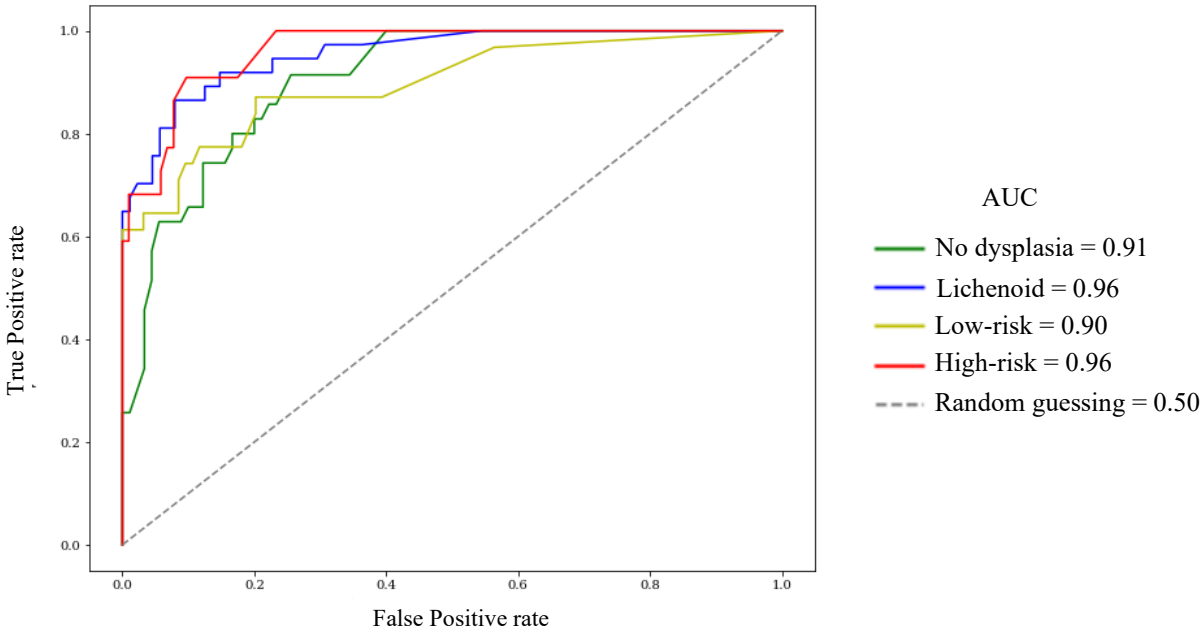


Figure 7.12. ROC curves for all diagnostic classes from the test results of the best ranked PyTorch fluorescein diagnostic triage CNN model

7.5. Comparison of all CNNs

The comparison of four CNN models across Acriflavine and Fluorescein staining techniques reveals nuanced performance differences across diagnostic categories. In the no dysplasia category, the MATLAB CNN with Fluorescein staining achieved the highest accuracy (88.80%) and demonstrated a well-balanced performance with sensitivity of 0.83 and specificity of 0.91 (Table 7.1.).

Table 7.1. Comparison of test performance of CNNs across both contrast agent datasets and development frameworks

Diagnostic category	**Metric**	**Acriflavine**		**Fluorescein**	
		MATLAB CNN	**PyTorch CNN**	**MATLAB CNN**	**PyTorch CNN**
No dysplasia	Accuracy	86.26%	54.20%	88.80%	86.40%
	Sensitivity	0.87	0.03	0.83	0.77
	Specificity	0.86	0.90	0.91	0.90
	Precision	0.81	0.17	0.78	0.75
	F1score	0.84	0.05	0.81	0.76
Lichenoid	Accuracy	90.08%	95.42%	84.80%	91.20%
	Sensitivity	0.78	0.90	0.76	0.92
	Specificity	0.93	0.97	0.89	0.91
	Precision	0.75	0.92	0.74	0.81
	F1score	0.76	0.91	0.75	0.86
Low-risk	Accuracy	87.79%	94.66%	87.20%	89.60%
	Sensitivity	0.71	0.87	0.65	0.74
	Specificity	0.96	0.97	0.95	0.95
	Precision	0.88	0.92	0.80	0.82
	F1score	0.79	0.89	0.71	0.78
High-risk	Accuracy	97.71%	51.91%	94.40%	94.40%
	Sensitivity	1.00	0.12	0.91	0.77
	Specificity	0.98	0.55	0.95	0.98
	Precision	0.74	0.02	0.80	0.89
	F1score	0.85	0.03	0.85	0.83

The Acriflavine MATLAB CNN followed closely with 86.26% accuracy and the highest sensitivity (0.87) in this category (Table 7.1.). For lichenoid cases, the PyTorch CNN with Acriflavine staining excelled, reaching 95.42% accuracy, 0.90 sensitivity, and an impressive 0.91 F1-score, outperforming all other models (Table 7.1.). In the low-risk category, the acriflavine PyTorch CNN model again led with 94.66% accuracy, 0.87 sensitivity, and 0.89 F1-score, showing strong discriminative power for this challenging category (Table 7.1.).

Table 7.2. All CNN models across both contrast agent datasets and development frameworks ranked based on test performance

Diagnostic category	**Metric**	**Acriflavine**		**Fluorescein**	
		MATLAB CNN	**PyTorch CNN**	**MATLAB CNN**	**PyTorch CNN**
No dysplasia	Accuracy	3	4	1	2
	Sensitivity	1	4	2	3
	Specificity	4	2	1	2
	Precision	1	4	2	3
	F1score	1	4	2	3
Lichenoid	Accuracy	3	1	4	2
	Sensitivity	3	2	4	1
	Specificity	2	1	4	3
	Precision	3	1	4	2
	F1score	3	1	4	2
Low-risk	Accuracy	3	1	4	2
	Sensitivity	3	1	4	2
	Specificity	2	1	4	3
	Precision	2	1	4	3
	F1score	2	1	4	3
High-risk	Accuracy	1	4	2	2
	Sensitivity	1	4	2	3
	Specificity	2	4	3	1
	Precision	3	4	2	1
	F1score	2	4	1	3
Aggregate rank		45	49	58	46
Final rank		1	3	4	2

The acriflavine MATLAB diagnostic triage CNN model performed the best across all models (Table 7.2.).

Table 7.3. Test confusion matrix for the best performing diagnostic triage CNN (Acriflavine - MATLAB) on in vivo confocal micrographs

		Predicted			
		No dysplasia	**Lichenoid**	**Low-risk**	**High-risk**
Actual	**No dysplasia**	94	8	5	1
	Lichenoid	9	42	3	0
	Low-risk	13	6	59	5
	High-risk	0	0	0	17

The model excelled at detecting high-grade OED & OSCC, which is crucial given the severity of these conditions (Table 7.3.). The high-risk category revealed the most striking results, with the acriflavine MATLAB CNN model achieving perfect sensitivity (1.00), the highest accuracy (97.71%), and an F1-score of 0.85 (Table 7.1.). This performance in the high-risk category is particularly crucial given the clinical importance of identifying potentially malignant cases.

Table 7.4. Test performance of the best diagnostic triage CNN models on human confocal micrographs across both contrast agents

		Cases correctly identified (sensitivity)	**Other categories wrongly predicted as this category** (1-specificity)	**Predictions of this category that were incorrect** (1 - precision)	**Cases misclassified** (1-sensitivity)
Low-grade OED	Acriflavine (MATLAB)	71.08%	4.47%	11.94%	28.92%
	Fluorescein (PyTorch)	74%	5%	18%	26%
High-grade OED & OSCC	Acriflavine (MATLAB)	100%	2.45%	26.09%	0%
	Fluorescein (PyTorch)	77%	2%	11%	23%

The Fluorescein PyTorch model performed better at identifying low-grade OED images (sensitivity: 74% vs 71.08%) (Table 7.4.). While for the high-grade OED and OSCC (high-risk) category the acriflavine model was superior as it correctly identified all the images. While 26.09% of its 'high-risk' predictions were unnecessary only 2.45% of other lesions images were incorrectly identified as being high-grade OED and OSCC making the acriflavine MATLAB CNN the best performing diagnostic triage model (Table 7.4.).

7.6. Discussion

The results in this study highlight the variation in performance of ML models between CNN model architectures, staining techniques, and implementation platforms in oral cancer detection.

The fluorescein CNNs demonstrated more consistent performance between MATLAB and PyTorch implementations, with the PyTorch version showing balanced performance across categories (Table 7.1.). This consistency could be advantageous in clinical settings where reliability across different computational environments may be crucial across different availabilities in computational resources. The fluorescein models' performance, while not topping any single category, showed fewer extreme variations, potentially offering more predictable results across diverse cases.

There was a clear difference in test performance between the acriflavine models in MATLAB and PyTorch implementations (Table 7.2.; ranking 1st and 3rd respectively out of 4), which raises important questions about the impact of implementation details on model performance. This discrepancy suggests that factors such as data preprocessing, model initialization, or optimization techniques between the MATLAB and PyTorch development environments, significantly influence outcomes emphasizing the need for standardized protocols in model development and evaluation.

The acriflavine-based MATLAB diagnostic triage CNN model outperformed all other models in this study. It was particularly effective in detecting high-grade oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC), conditions where early and accurate detection is critical due to their severity. The model achieved 100% sensitivity for these high-risk lesions, meaning it did not miss any cases, and it maintained a very low false positive rate of 2.4%.

For low-grade OED, the model’s performance was good but less robust. It correctly identified 71.1% of cases, missing over 1 in 4 cases Interestingly, the false discovery rate was higher for high-grade OED and OSCC (26.1%) than for low-grade OED (11.9%), suggesting the model tends to err on the side of caution, sometimes flagging benign cases as high-risk. If applied for screening this could lead to some unnecessary urgent biopsies but ensures that no high-risk cases are missed.

This cautious but thorough approach gains relevance when compared to real-world clinical decision-making. In a large population-based study by Chaturvedi et al. (2020) assessing biopsy decisions for oral leukoplakia in Northern California (USA), only 59.6% of the decisions were correct. This low sensitivity meant 40.4% of cancerous lesions were missed, and 94.9% of biopsied lesions turned out to be non-cancerous (Chaturvedi et al., 2020). These findings highlight the potential value of decision-

support tools like the CNN model developed in our study, which could help reduce missed diagnoses while supporting more consistent triage decisions.

A comparable non-invasive approach to confocal microscopy is the Visually Enhanced Lesion Scope (VELScope), a handheld device that enables direct visualization of tissue fluorescence changes using blue excitation light (400–460 nm). It has been investigated as a tool for oral cancer screening (Yeladandi, Sundaram, Muthukumaran, Umamaheswari, & Dhanya, 2024). In a study of 118 lesions, VELScope demonstrated a sensitivity of 30% for detecting oral potentially malignant disorders (OPMDs), compared to 25% for clinical oral examination alone (Farah, McIntosh, Georgiou, & McCullough, 2012). When both methods were combined, sensitivity improved to 46% (Farah et al., 2012). However, this is still substantially lower than the 71.1% sensitivity achieved in the current study using confocal microscopy combined with a CNN model for detecting dysplasia. These findings highlight the improved diagnostic performance of the confocal CNN-based approach over existing non-invasive tools like VELScope.

A perfect AUC of 1 for high-risk images obtained in this study draws caution to the possibility of this CNN is overfitting to that class. The high-risk class also had the fewest images indicating a high probability of the CNN memorising the features of this class rather than learning them for generalisation. Overfitting occurs when a model performs well on its training data but poorly on new, unseen data, indicating it has learned noise rather than general patterns. While internal validation methods like train-test splits help, a true test of overfitting requires external validation. This means testing the model on an entirely new dataset from a different sample or population to genuinely assess its real-world generalisability. This involves assessing the model's performance on an entirely new and independent dataset from a different source. This new dataset should ideally come from different populations, varying by geography, demographics (age, ethnicity, gender), clinical characteristics, or even time periods, to truly confirm the model's real-world applicability and robustness.

Although the best-performing CNN model demonstrated promising results, it requires further refinement before it can be considered clinically applicable. A key step toward improvement would be the use of a larger, more balanced dataset that includes comparable sample sizes across all diagnostic categories, enabling the model to better learn representative features. The transformative impact of large-scale datasets on CNN performance was notably illustrated by the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where the availability of 14 million labelled images significantly advanced the capabilities of deep learning models (Deng, Dong, Socher, Li, Kai, et al., 2009). This foundational dataset contributed to the development of high-performing architectures such as Inception_V3, which was employed in this study (Szegedy et al., 2016).

Inception_V3 was expected to adapt and fine-tune its feature extraction capabilities to the specific characteristics of this limited custom dataset in the current study.

Remarkably, the model achieved strong performance despite being trained on fewer than 2000 confocal microscopy images. This was a particularly challenging scenario due to severe class imbalance, with high-grade OED and OSCC samples representing the smallest categories.

A prospective trial designed where a larger fluorescence in vivo confocal microscopy dataset collected with the specific aim of AI analysis with a balanced dataset across different OED and OSCC samples would be ideal. Additionally, representing a variety of demographics by conducting a multi-centre trial across different sites globally would enhance the generalisability of the models developed.

These limitations in the training and validation of the best performing CNN model from the present study raises the question of clinical readiness. While the algorithm can be integrated with a cloud-based program within a software as a service model, its clinical reliability can be questionable at best based due to its limited testing. External validation via a multi-centre trial would be the next step before software integration can be explored. A recent review of artificial intelligence algorithms for automated interpretation of screening mammography identified 13 external validation studies which evaluated CNN models (Anderson et al., 2022). While most studies identified in this review had a high or unclear risk of bias (69%) due to sampling designs, it was a sign of improvement of external testing of CNNs in that field (Anderson et al., 2022). This step forward can be replicated in dentistry and specifically oral cancer detection.

If and when that is a possibility, the recommendation for using the CNN developed in the present study could be incorporate into a software pipeline in tandem with the quality filtering CNN developed in Chapter 4. Quality in-quality out is a popular philosophy in machine learning, and it aligns with the gradient descent principles machines use to learn (LeCun et al., 2015a).

CNNs can provide predictions with a high level of confidence and there is a possibility of these models to be confidently incorrect. This dynamic becomes particularly problematic when AI systems encounter inputs that deviate from their training data, known as distributional shift (Davis, Lasko, Chen, Siew, & Matheny, 2017). Rather than indicating uncertainty, many AI models continue to issue confident predictions. Without mechanisms to communicate confidence or uncertainty, these systems can produce misleading outputs that are not easily observed, especially in fast-paced or resource-constrained clinical environments (Davis et al., 2017). A key concern is ‘automation complacency’, where clinicians overly rely on AI outputs, especially when these systems have a history of high accuracy. In such contexts, users may stop actively scrutinizing AI recommendations, assuming the system to be infallible. This behaviour is exacerbated when clinicians have busy workflows with high cognitive load and fatigue, making them less likely to question AI decisions. This is particularly likely when those AI decisions align with the clinicians’ initial diagnostic impressions, which can be described as confirmation bias (Challen et al., 2019).

Parasuraman & Manzey's work in this area revealed that when clinicians juggle multiple tasks, attention is naturally reallocated away from monitoring automation, especially if the AI is perceived as trustworthy. Notably, both novice and expert users are susceptible to these effects, and training alone is often insufficient to mitigate them. Automation bias, in particular, results in omission errors (failing to act because the AI does not flag an issue) and commission errors (acting on incorrect AI advice), both of which can lead to diagnostic failures (Parasuraman & Manzey, 2010). While improving AI model performance is one solution for combating these issues, AI literacy for clinicians would be vital to preserve critical thinking in clinical decision-making.

Incorporating this technology into clinical practice, based on its test classification speed each image captured in vivo using a fluorescence confocal microscope could undergo diagnostic triage in less than $1/10^{th}$ of a second. Entire lesions across multiple intra-oral sites can be imaged and triaged within a couple of minutes assuming the imager can capture diagnostic quality images. Regardless of its potential, this technology does not threaten to eliminate the standard of care biopsy with histopathological assessment. However, it will allow for a much higher rate of accurate sampling across multiple lesions in a painless non-invasive manner within a matter of minutes. This can exponentially improve diagnostic precision and help to monitor patients efficiently across multiple visits.

8. DIAGNOSTIC TRIAGE DEEP LEARNING MODELS IN A MURINE MODEL OF ORAL CARCINOGENESIS

8.1. Introduction

Detection of oral cancer in its early stages is crucial for enhancing survival rates, reducing morbidity, reducing treatment duration, improving psychological outcomes, and enhancing overall quality of life (Saka-Herrán, Jané-Salas, Mari-Roig, Estrugo-Devesa, & López-López, 2021). Currently, the majority of oral cancer cases are diagnosed in advanced stages contributing to 5-year survival rate of only 50% (Grafton-Clarke, Chen, & Wilcock, 2019; Güneri & Epstein, 2014). The duration between the onset of initial symptoms, which are often mild, diagnosis and specialist referral plays a pivotal role in influencing disease staging and ultimately patient prognosis (Seoane et al., 2016).

At the microscopic level oral epithelial dysplasia (OED) refers to a range of architectural and cytologic alterations impacting the epithelial lining of the oral mucosa, arising from the accumulation of genetic alterations (EI-Naggar, 2017). This fertile environment of chromosomal instability is associated with increased risk for carcinoma (Odell et al., 2021a). While not all cases of OED advance to oral squamous cell carcinoma (OSCC), the heightened risk of progressing to malignancy can influence clinical management of dysplastic oral lesions (Odell et al., 2021a). Therefore, earlier detection of oral lesion dysplasia could aid in monitoring these suspicious lesions for more timely therapeutic intervention.

Mouse oral cancer in vivo models are dynamic systems that most closely replicate the progression, histopathology, and molecular characteristics of human oral cancers. They offer the ability to create controlled and reproducible environments, allowing researchers to introduce specific genetic mutations or environmental exposures to study tumour initiation, progression, and therapeutic responses with high precision (Luo, Young, Zhou, & Wang, 2018). The current standard of care in mouse oral cancer models involves clinical staging followed by euthanasia and histopathological analysis, which limits long term assessment and increases cost. Developing non-invasive in vivo visualization methods would enable real-time monitoring of tumour progression, reduce the number of animals required, increase experimental yield, and provide critical insights into the timing and effectiveness of therapeutic interventions (Luo et al., 2018).

Several chemical carcinogens coupled with conventional histopathology have been utilized in animal models to study the early microscopic changes associated with oral cancer (Sagheer et al., 2021). Of these carcinogens, 4-nitroquinoline-1-oxide (4-NQO) is the most commonly utilized (Kanojia & Vaidya, 2006). 4NQO initiates random mutations by irreversibly reacting with the nucleophilic segment of DNA, preferentially forming DNA adducts at guanine residues. These adducts result in guanine-to-pyrimidine substitutions, contributing to mutations in carcinogenic genes (Downes et

al., 2014). Alongside DNA damage, 4NQO produces reactive oxygen species (ROS), causing additional harm to DNA, proteins, and lipids.

This multifaceted damage process significantly contributes to the progression of tumours (Koike, Uchiyama, Arimoto-Kobayashi, Okamoto, & Negishi, 2018). Similar to the human disease, the transition from hyperplasia to dysplasia indicating an onset of OED is difficult to pinpoint. Detection of early signs of OED in mice could facilitate early optimal testing of novel therapeutics and ultimately earlier oral cancer patient diagnosis. Imaging OEDs and oral cancer with the help of non-invasive digital microscopy imaging in the form of confocal microscopy could provide an automated workflow for screening. Confocal microscopy has been explored in humans for early detection of oral potentially malignant disorders and oral cancer utilising several image analysis approaches with variable success (Ramani et al., 2023).

Convolutional neural networks (CNN) are the most effective deep learning models for addressing computer vision challenges such as facial recognition algorithms, guidance systems for autonomous vehicles, product identification in self-service supermarkets (LeCun et al., 2015a). Functioning as feed-forward networks, CNNs excel in feature extraction from image data through pixel convolution, drawing inspiration from biological visual perception neurons (LeCun et al., 2015a). Various CNN architectures have undergone rigorous testing in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), with the Inception_v3 model showcasing outstanding performance (2015 edition - top-5 error rate = 3.58%). CNN models such as Inception_v3 may be repurposed for image classification tasks beyond their original scope through transfer learning (Weiss, Khoshgoftaar, & Wang, 2016).

The universal deep learning framework, PyTorch simplifies implementation of CNNs utilizing Python programming language to manage the complexities of deep learning with intuitive surface-level APIs for model development (Paszke et al., 2019). In our study we also used PyTorch’s ability to generate computational graphics during model training, with the "autograd" module to automate gradient computation (Paszke et al., 2017a). And integration with GPU acceleration to enhance the efficiency of training our large large-scale CNNs (Karimi, Dou, Warfield, & Gholipour, 2020). We also used several tools to integrate trained PyTorch CNN models into software or cloud computing applications (Paszke et al., 2019). Another development framework used for machine learning models is the MATLAB (MathWorks, USA) package called Deep Network Designer (Kim, 2017). In the present study CNNs developed using MATLAB and PyTorch from in vivo lesions captured confocal micrographs from mice during a 4NQO model of oral cancer were used to assess their ability to differentiate between low-grade and high-grade cancerous epithelial dysplasia.

Aim: To develop and evaluate the performance of convolutional neural network (CNN) models for the early detection and classification of oral epithelial dysplasia (OED) and

oral cancer using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

8.2. Methods

This study involved developing and evaluating deep learning convolutional neural network (CNN) models for the classification of oral epithelial dysplasia (OED) and oral cancer using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model. The study used a mouse model with 4-NQO-induced oral cavity lesions, following strict ethical guidelines.

All experiments were approved by The Walter and Eliza Hall Institute Animal Ethics Committee in accordance with relevant guidelines and regulations. Mice were housed at The Walter and Eliza Hall Institute animal facility under specific pathogen-free, temperature- and humidity-controlled conditions with a 12 h light/dark cycle and ad libitum feeding. Two types of genetically modified mice were used: those without functional c-REL alleles (c-Rel-/-) and those lacking TNF (Tnf-/-), both backcrossed onto a C57BL/6 background for >10 generations. Control wild-type (C57BL/6) mice were housed in adjacent boxes within the same room.

Oral cavity lesions were induced in 14-week-old mice by treating them with regular drinking water ± 100μg/ml 4-NQO for 14 weeks. Following treatment, mice were given regular drinking water ad libitum until necropsy 3-12 weeks post-treatment cessation. Confocal imaging was carried out at weeks 14,16,18,22,24, and 26 which marked when the respective mice were euthanised (Figure 3.16. from Chapter 3). This facilitated the capture of different stages of the dysplastic process. Confocal images were captured using the hand-held point scanning confocal microscope ViewnVivo FIVE2 (Optiscan Imaging, Australia), equipped with a 3.5 mm diameter x 66 mm long hand-held probe.

The probe was directly applied to regions of interest within the oral cavity which included lesions developed at the epithelial surface of the tongue (dorsal/ventral) and the buccal mucosa for image acquisition. The imaging was carried out by a dentist with expertise in oral mucosal disease (A/Prof. Antonio Celentano) from the Melbourne Dental School, University of Melbourne (Victoria, Australia).

The procedure for imaging the oral cavity in mice involves several steps (Protocol 6, Chapter 3):

1. Mice were sedated using a combination of medetomidine, midazolam, and fentanyl
2. The oral cavity was then clinically assessed and photographed
3. The cavity was cleansed using saline-soaked cotton swabs to remove debris
4. Micro-applicator brushes soaked in 10–15 μL of 0.1% acriflavine contrast agent solution were used to gently paint the oral surfaces and left for one minute before excess was removed by gently wiping with cotton swabs

5. Confocal microscopy imaging was performed using a ViewnVivo FIVE2 microscope, with the probe in direct contact with the oral tissue

This was followed by biopsy and histopathology scoring by an independent pathologist based on the binary grading of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) for establishing the ground truth dysplasia grade (Kujan et al., 2006).

The diagnostic categories for confocal microscopy CNN analysis based on the histopathology reference (Figure 8.1.):

1. no dysplasia
2. low-grade dysplasia
3. high-grade dysplasia and OSCC

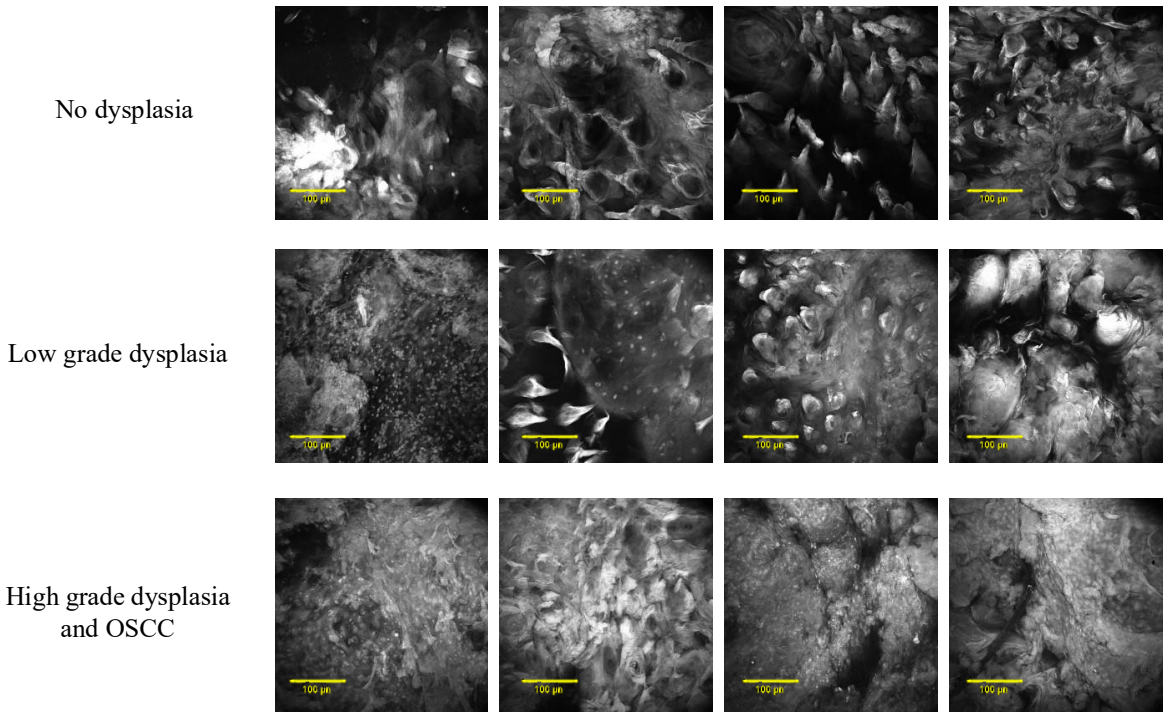


Figure 8.1. Examples of acriflavine confocal micrographs across the diagnostic categories, displaying the variation in appearance of images across and within the diagnostic categories

For CNN development, confocal micrographs were labelled based on the binary dysplasia pathology scoring of H&E samples (Figure 2). Images were modified to RGB colour format and down-sampled to 299 x 299 pixels. The dataset was split into training (80%) and test (20%) sets.

CNNs were developed using MATLAB's Deep Network Designer application and the PyTorch framework and Python 3 programming language. The CNN architecture was a modified Inception_V3 with transfer learning. The MATLAB CNN model was

developed using the default set-up of 30 epochs and a learning rate of 0.001. The PyTorch CNN development involved hyperparameter optimization using a grid search approach, exploring combinations of epochs (5,10,15) and learning rates (0.001, 0.01, 0.1). K-fold cross-validation (k = 5) was employed, resulting in 45 CNN models being trained and tested within the PyTorch development environment. Results of CNNs trained and tested using both development approaches (MATLAB and PyTorch) were compared.

CNN performance was assessed using multiple evaluation metrics:

1. **Accuracy:** The proportion of correctly classified instances out of the total instances.
2. **Sensitivity (Recall):** The proportion of actual positives correctly identified by the model.
3. **Specificity:** The proportion of actual negatives correctly identified by the model.
4. **Precision:** The proportion of predicted positives that are actually positive.
5. **F1-score:** The harmonic mean of precision and recall, balancing both metrics.
6. **Received operator characteristic (ROC) Curve:** A plot of the true positive rate (sensitivity) against the false positive rate across different thresholds.
7. **AUROC (Area Under the ROC Curve):** A single value summarizing the ROC curve, indicating the model's ability to distinguish between classes.

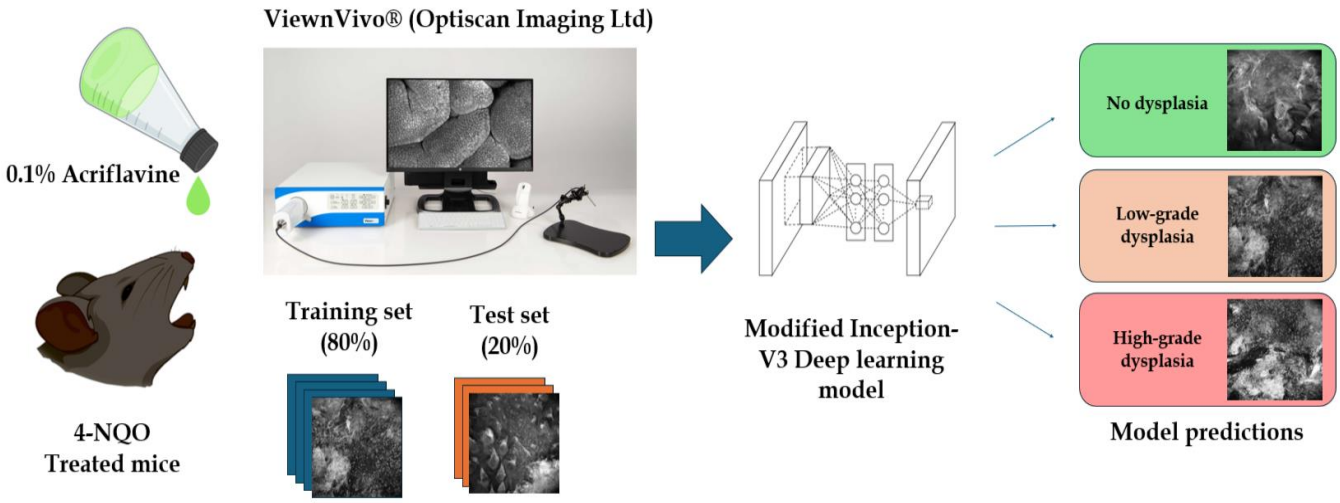


Figure 8.2. Depiction of the CNN image analysis pipeline for murine acriflavine images

The PyTorch models were ranked based on aggregated performance across metrics and classes. These were compared with the MATLAB CNN. All Python and PyTorch code

used in this study are available online at the provided GitHub repository (Appendix 2). Statistical analyses were performed using the Python sci-kit learn library.

8.3. Data distribution

The study analysed 45 unique CNN models being trained in PyTorch across all hyperparameter combinations in the grid search optimisation process. These models were then evaluated using a 5-fold cross-validation strategy, and the best-performing model for each approach was identified based on an aggregation of performance metric ranks. The image distribution across the diagnostic categories depicted in Table 8.1.

Table 8.1. Training and test dataset split for all classes in the murine acriflavine dataset

	Training	**Test**	**Total**
Class 1: No dysplasia	1819	449	2268
Class 2: Low-grade dysplasia	1262	258	1520
Class 3: High-grade dysplasia and Oral squamous cell carcinoma	859	239	1098
Total	3940	946	4886

8.4. MATLAB murine diagnostic CNN

The MATLAB murine diagnostic CNN model was trained on 3940 images and tested on 946 previously unseen images across all 3 diagnostic categories (Table 8.2.). The MATLAB CNN had a moderate accuracy at detecting no dysplasia (62.68%) and low-grade OED lesions (62.90%) with higher accuracy with high-grade OED & OSCC lesions (77.59%) (Table 8.2.).

Table 8.2. Test confusion matrix for the MATLAB murine CNN model

		Predicted		
		No dysplasia	Low-grade dysplasia	High-grade dysplasia and OSCC
Actual	No dysplasia	216	163	70
	Low-grade dysplasia	83	167	8
	High-grade dysplasia and OSCC	37	97	105

The sensitivity of identifying these lesions ranged from 0.44 (high-grade OED &OSCC) to 0.65 (low-grade OED) and the specificity was on the higher end ranging from 0.62 (low-grade OED) to 0.89 (high-grade OED & OSCC) (Table 8.3.). The precision of this model was poor to moderate with values ranging from 0.39 (low-grade OED) to 0.64 (no dysplasia) (Table 8.3.). The overall F1 scores of this model were also moderate ranging from 0.49-0.55 (Table 8.3.).

Table 8.3. Test results of the MATLAB murine diagnostic CNN

Class	Accuracy (%)	Sensitivity (recall)	Specificity	Precision	F1 score
No dysplasia	62.68	0.48	0.76	0.64	0.55
Low-grade OED	62.90	0.65	0.62	0.39	0.49
High-grade OED & OSCC	77.59	0.44	0.89	0.57	0.50

8.5. PyTorch murine diagnostic CNN

The PyTorch murine diagnostic CNN models were all trained on 3940 images and tested on 946 previously unseen images across all 3 diagnostic categories. These model performance scores were averaged across all cross-validation folds to obtain test results for each hyperparameter combinations (Table 8.4.).

Table 8.4. Test results models averaged across all epochs and learning rate combinations

Number of epochs		**5**	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
Learning rate		**0.001**	**0.001**	**0.001**	**0.01**	**0.01**	**0.01**	**0.1**	**0.1**	**0.1**
No dysplasia	**Accuracy (%)**	58.54	62.11	61.29	61.04	60.74	60.85	60.95	66.13	65.96
	Sensitivity	0.75	0.61	0.55	0.56	0.51	0.54	0.61	0.65	0.64
	Specificity	0.43	0.63	0.67	0.66	0.70	0.67	0.61	0.67	0.68
	Precision	0.55	0.60	0.60	0.60	0.60	0.60	0.59	0.64	0.64
	F1 score	0.63	0.60	0.58	0.57	0.55	0.57	0.59	0.65	0.64
Low-grade OED	**Accuracy (%)**	64.63	63.81	63.11	62.81	65.22	65.73	64.74	65.75	67.51
	Sensitivity	0.42	0.56	0.59	0.59	0.58	0.57	0.55	0.57	0.55
	Specificity	0.73	0.67	0.65	0.64	0.68	0.69	0.68	0.69	0.72
	Precision	0.37	0.39	0.38	0.38	0.41	0.41	0.41	0.41	0.43
	F1 score	0.39	0.46	0.47	0.46	0.48	0.48	0.46	0.48	0.48
High-grade OED & OSCC	**Accuracy (%)**	76.62	80.04	78.52	75.94	73.66	75.54	73.21	73.36	73.81
	Sensitivity	0.11	0.35	0.36	0.30	0.39	0.39	0.22	0.24	0.33
	Specificity	0.99	0.95	0.93	0.92	0.85	0.88	0.90	0.90	0.88
	Precision	0.84	0.72	0.63	0.55	0.49	0.52	0.47	0.45	0.48
	F1 score	0.18	0.47	0.46	0.38	0.42	0.44	0.30	0.31	0.39

For ‘no dysplasia’ the accuracy ranged from 58.54% to 66.13%, with relatively consistent precision (0.55–0.64) and F1 scores (0.57–0.65). Sensitivity varied significantly (0.55–0.75), and specificity was generally low to moderate (0.43–0.70) (Table 8.4.). Low-grade dysplasia results showed a higher accuracy than no dysplasia ranging from 63.11% to 67.51% (Table 8.4.). However, precision was low (0.37–0.43), and sensitivity remained modest (0.42–0.59), resulting in F1 scores of 0.39–0.48 (Table 8.4.). The lower precision and recall suggest that the model struggles to effectively differentiate between true positives and false positives of mild dysplasia images. High-grade dysplasia and OSCC achieved the highest accuracy among all groups (73.21%–80.04%) and the best specificity (0.85–0.99) (Table 8.4.). However, sensitivity was consistently low (0.11–0.39), and F1 scores were moderate (0.18–0.47) (Table 8.4.). The strong specificity suggests the model is excellent at identifying negative cases for the high-grade OED & OSCC group, but its poor sensitivity highlights difficulty in identifying these cases.

The model with the best performance rank, trained in fold 5 with 10 epochs and a learning rate of 0.1 was selected to be the PyTorch murine diagnostic CNN model (Table 8.5.).

Table 8.5. Test confusion matrix for the PyTorch murine CNN model

		Predicted		
		No dysplasia	Low-grade dysplasia	High-grade dysplasia and OSCC
Actual	No dysplasia	294	172	113
	Low-grade dysplasia	128	154	88
	High-grade dysplasia and OSCC	155	174	81

This model showed a moderately high sensitivity (0.65), specificity (0.69), precision (0.66) and F1 score (0.66) in identifying normal & hyperplasia micrographs (Table 8.6.). Its performance for all metrics except specificity dipped for mild, moderate and severe dysplasia & carcinoma (Table 8.6.). The model’s low sensitivity at identifying moderate, severe, and carcinoma images (0.34) and high specificity (0.90) indicated its ability to identify lesions that were not severe (Table 8.6.). Its moderate precision (0.53) at identifying these high severity lesions indicated its predictions of these lesions was correct about half the time (Table 8.6.).

Table 8.6. Test results of the best ranking PyTorch model (Fold 5, Epochs 10, Learning rate 0.1)

Class	Accuracy (%)	Sensitivity (recall)	Specificity	Precision	F1 score
No dysplasia	67.44	0.66	0.69	0.66	0.66
Low-grade OED	68.82	0.60	0.72	0.45	0.51
High-grade OED & OSCC	75.58	0.34	0.90	0.53	0.41

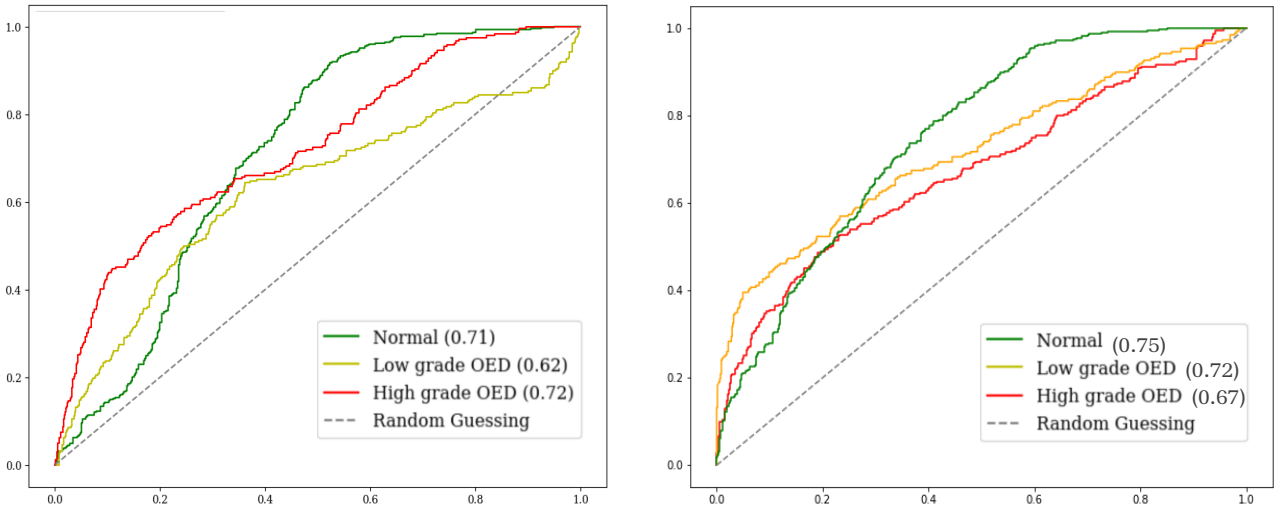
8.6. Comparison of MATLAB and PyTorch murine diagnostic CNNs

Both the MATLAB and PyTorch models had a very similar performance on the test dataset. The MATLAB model marginally outperformed the PyTorch model in terms of sensitivity of detecting OED and OSCC (Tables 8.5.). The PyTorch model performed better for the 'No dysplasia' and 'Low-grade OED' classes across most metrics while the MATLAB model performed better for the 'High-grade OED & OSCC' class, particularly in sensitivity, and F1 score. Both models showed the highest accuracy for the 'High-grade OED & OSCC' class (Table 8.3. & 8.4.). The PyTorch model showed more balanced performance across classes, while the MATLAB model had more variability between classes. Both models struggled with sensitivity for the 'High-grade OED & OSCC' class, indicating difficulty in correctly identifying positive cases in this category (Table 8.5.)

Table 8.7. Performance of the murine diagnostic CNNs on detection of OED and OSCC

		Cases correctly identified (sensitivity)	**Other categories wrongly predicted as this category** (1-specificity)	**Predictions of this category that were incorrect** (1 - precision)	**Cases misclassified** (1-sensitivity)
Low-grade OED	**MATLAB**	64.73%	37.79%	60.89%	35.27%
	PyTorch	43.93%	11.03%	42.62%	56.07%
High-grade OED & OSCC	**MATLAB**	59.69%	27.76%	55.36%	40.31%
	PyTorch	33.89%	10.33%	47.40%	66.11%

The MATLAB CNN identified only 64.73% of low-grade OED images while the misclassifying 40.31% of high-grade OED & OSCC images (Table 8.5.). The PyTorch CNN misclassified 56.07% of low-grade OED and 66.11% of high-grade OED & OSCC cases.



MATLAB murine diagnostic CNN

PyTorch murine diagnostic CNN

Figure 8.3. Receiver Operator Curves (ROC) and area under the curve (AUC) for the best MATLAB and PyTorch murine diagnostic CNNs

Both models show AUROC scores above 0.5 for all classes, indicating that they perform better than random chance (0.5) in distinguishing between classes (Figure 8.3.). The PyTorch model shows more consistent performance across classes (range: 0.67-0.75) compared to the MATLAB model (range: 0.62-0.72) (Figure 8.5.). While the scores are above 0.5, there is room for improvement in both models.

8.7. Discussion

The murine oral cancer model was selected to test the CNN diagnostic triage approach used in the human confocal microscopy dataset, as the success of non-invasive AI diagnostics could offer some key benefits. First, AI enables faster hypothesis generation and validation by rapidly analysing large complex datasets generated in a lab, identifying patterns, and predicting outcomes that might otherwise take months to uncover through traditional methods. When connected with non-invasive tools such as probe based in vivo confocal endomicroscopy, it could reduce the need for sacrificing animals in certain experiments, enabling longitudinal studies of disease progression and novel therapeutic response studies while upholding more ethical research standards.

This study involved developing CNNs for the diagnosis of OED and OSCC using the MATLAB and PyTorch deep learning frameworks and comparing the resulting models. Both models misclassified majority of the OED & OSCC cases and were therefore considered unsuitable for clinical application (Table 8.5.). This drop in performance between the human image dataset (Chapter 7) and the murine dataset despite being trained and tested on by the same CNN models is potentially linked to the nature of the images themselves.

The murine confocal micrographs had a large variation in appearance ranging from superficial to deep cross sections of the tissue being imaged. Several confocal micrographs of the tongues of mice having lesions across all 3 diagnostic categories appeared to focus on the papillae on the surface of the tongue. This superficial field of view might provide some value but information pertaining to the deeper layers of the epithelium which showcase signs of oral epithelial dysplasia are ignored. This issue was compounded by the variation in appearance of micrographs within and between the diagnostic categories potentially leading to feature confusion in the CNNs (Liu & Zhang, 2019). Within the same tissue type (e.g. healthy epithelial cells) confocal microscopy images can differ due to varying topical contrast agent penetration, different field of view and altered focal plane due to varying penetration depths of incident laser from the confocal microscopy probe. A quality filtering approach as was conducted for human images was avoided for the murine dataset due to its carefully curated in-lab imaging which could have potentially impacted the results.

The variation in appearance could be due to the challenges with in vivo imaging in mice. The oral cavity of a mouse is very small. The 3.5mm diameter probe of the ViewnVivo system (Optiscan Imaging, Australia) while still small for use in human patients is quite large in the context of a murine mouth. This makes it difficult to manoeuvre and orient the probe to reach all areas where lesions form in the oral cavity (Celentano, Rickard, Low, Silke, Mohammed, Moslemi, Ramani, Franca, Reiner, & McCullough, 2025). Passive factors such as saliva and mucus can obscure the field of

view. Even under anaesthesia, subtle movements related to breathing and swallowing can introduce artifacts or shift the tissue, making precise imaging challenging. Additionally, 4NQO can induce a variety of lesions on the oral mucosa. These involve cauliflower-shaped that make imaging the portions of the oral epithelium close to the basement membrane challenging (Celentano, Rickard, Low, Silke, Mohammed, Moslemi, Ramani, Franca, Reiner, & McCullough, 2025).

Another potential reason for the poor performance could be the low inter-class variation in appearance of the confocal micrographs. Distinct tissue types among the selected diagnostic categories may appear visually similar on fluorescence confocal micrographs in the 4NQO oral carcinogenesis model (Figure 8.1.). The CNN may extract overlapping features for these classes, making it difficult to distinguish between them. The CNNs being ‘black boxes’ with millions of parameters prevent an interpretation of the reasoning and feature selection (LeCun et al., 2015a). Ultimately the CNNs developed for the murine fluorescence confocal microscopy dataset showed very limited application on the murine oral carcinogenesis model due to misclassifying close to 1/3rd of OED and OSCC cases.

This CNN based approach of image analysis could be further improved in a controlled animal study by collecting a larger sample size of images from varied mice. The imbalance in samples across diagnostic categories could be better controlled to prevent the CNNs from developing class-specific biases. Other types of CNN architectures could be trialled on this murine confocal microscopy dataset to explore the potential identification of new features. To address the ‘black box’ issue advances in explainable AI could be applied to the CNNs tested in order to fine tune them specifically for a better overall performance.

The integration of in vivo confocal microscopy with CNN models presents a formidable strategy for oral cancer detection in murine models, even when initial results are suboptimal. Confocal microscopy's primary advantage lies in its ability to provide real-time, high-resolution "optical biopsies" of the mouse oral cavity (Rangrez, Bussau, Ifrit, & Delaney, 2021). This non-invasive approach allows for the repeated imaging of the same lesions over time, which is very useful for longitudinal studies tracking disease progression or therapeutic responses within individual animals, thereby reducing the need for large animal cohorts. Additionally, CNNs can be iteratively refined and improved with more data and advanced architectures, meaning that even moderately poor initial results serve as valuable benchmarks for future development (LeCun et al., 2015a).

While the CNNs in this study performed poorly at triaging the mice for potentially malignant and malignant oral lesions, these results are promising signs of using this technology in preventing animal sacrifice in novel cancer detection and therapy testing studies.

9. DISCUSSION & CONCLUSIONS

The primary goal of this dissertation was to develop models that could assist in the diagnosis of oral potentially malignant disorders and oral squamous cell carcinoma to eventually support clinical decision-making.

Given that dysplastic changes vary in their malignant potential, stratifying individuals into risk categories based on predicted probability estimates can provide a practical framework for guiding clinical management. Rather than assigning arbitrary classifications, our approach was grounded in the potential clinical implications associated with different dysplasia grades identified by histopathology, with the 'high risk' category corresponding to cases more likely to require intervention or further investigation. This categorisation of disease detection was an implicit assessment of how well the models distinguish between those with and without histopathologically significant disease.

This binary classification approach adapted from Kujan et al. (2006) stratifying patients based on predicted risk of malignance was designed to align with clinical actions, such as decisions regarding referral for biopsy, closer surveillance, or surgical intervention (Kujan et al., 2006). Importantly, the threshold used to differentiate high- from low-risk was selected based on the potential clinical consequences of false positives and false negatives. For example, failing to identify a patient with high-grade dysplasia (a false negative) may delay necessary treatment, whereas classifying a patient with benign or low-grade dysplasia as high-risk (a false positive) may lead to unnecessary anxiety or interventions.

These considerations highlight that model evaluation must go beyond overall accuracy metrics. While statistical measures such as accuracy, sensitivity, specificity, precision, F1 score and AUROC were used to gauge performance, we recognise that model utility is ultimately determined by how well these outputs support clinical decisions. Future work should incorporate formal decision-analytic approaches to better quantify the trade-offs between different types of diagnostic errors in this context.

9.1. Hypothesis 1: Systematic review

Aim: To summarise and evaluate the evidence on the utility and performance of confocal microscopy in the diagnosis of oral potentially malignant disorders and oral squamous cell carcinoma.

Null Hypothesis (H_0): That studies do not utilise confocal microscopy for diagnosing oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That studies utilise confocal microscopy for diagnosing oral potentially malignant disorders and oral squamous cell carcinoma.

The systematic review detailed in Chapter 2, notes the efficacy of analysing the images produced by the CLE technology on the diagnosis of OPMDs and OSCC using a variety of approaches.

The findings of this review confirmed that the existing literature on the use of confocal microscopy for detecting OSCC and OPMDs is highly heterogeneous, both in methodological approach and reporting standards. Variability in study designs, imaging modalities (e.g., reflectance vs fluorescence), diagnostic thresholds, and outcome measures made direct comparisons challenging and limited the ability to perform meaningful meta-analysis. This fragmentation underscores the need for standardized protocols and consensus on diagnostic criteria in future research.

While several studies reported encouraging diagnostic performance, the inconsistencies across the evidence base prevented definitive conclusions about the clinical reliability or generalisability of confocal microscopy in this context. These limitations justify the need for carefully designed prospective studies with unified methodologies to better assess its role as an adjunctive or alternative tool in the diagnosis and monitoring of OSCC and OPMDs.

Out of the 28 studies identified, 22 (78.6%) utilised qualitative approaches while only 6 (21.4%) employed a quantitative approach (Ramani et al., 2023). The qualitative studies focused on locating standard of care histopathology features of OPMDs and OSCC such as cellular and architectural disarray, keratin pearls, nucleus cytoplasmic ratios and connective tissue features in their confocal micrographs. The results of these qualitative methods ranged from 0.5-1 sensitivity and 0.79-1 specificity when examined by oral pathologists and surgeons with microscopy analysis skill and experience ranging from

novice to expert (Linxweiler et al., 2016b; Moore et al., 2016; Oetter et al., 2016; Shavlokhova, Sandhu, et al., 2021; Ulrich et al., 2011a).

However, the comparison of histopathology features found in cross sectional biopsy samples with en face oriented (parallel to tissue surface) confocal microscopy images has the potential to be unreliable. The appearance of the epithelial cell architecture, which is crucial for identifying landmarks in histopathology, can vary considerably when viewed from a perspective that is offset by 90 degrees or more.

Although several studies in the review reported high performance in oral cancer and OPMD detection using their respective methods, the studies with quantitative, and statistical approaches seemed to have the most potential for reproduction and consistent application. The 6 quantitative analysis studies described in the review described a variety of approaches. They analysed fluorescence intensity, cell border shape, and connective tissue blood vessel diameter to identify OSCC (Ramani et al., 2023). Some studies in the review highlighted the challenges of in vivo imaging with different confocal microscopy devices. Targeting of lesion with the tip of the endoscope/confocal microscope probe could be difficult, as a stabilization of the handheld confocal microscope has been a challenge (Haxel et al., 2010).

Other challenges involve artifacts due to breathing movements in the subject or mucous/blood accumulation on the optical probe of the confocal microscope (Pogorzelski, Hanenkamp, Goetz, Kiesslich, & Gosepath, 2012b). Despite the quantitative approaches presented, there was a lack of unique standardised criteria for confocal microscopy images in the identification of OPMDs and OSCC. This systematic review highlighted the need for uniform, reproducible quantitative confocal microscopy image analysis methods for the accurate detection of OPMDs and OSCC.

While confocal microscopy shows promise for detecting OSCC and OPMDs, its current clinical reliability is limited by inconsistencies in methodology and a lack of standardised, quantitative diagnostic criteria. This highlights the need for standard protocols to enable its effective use as a diagnostic adjunct in clinical practice.

Since confocal microscopy was used for the detection of OPMDs and OSCC the null hypothesis was rejected.

9.2. Hypothesis 2: Quality filtering micrographs

Aim: To develop a CNN model that can identify diagnostic quality fluorescence in vivo confocal micrographs for downstream diagnostic triage by filtering out poor quality data.

Null Hypothesis (H_0): That a CNN model cannot accurately and rapidly identify fluorescence in vivo confocal micrographs of high diagnostic quality.

Alternative Hypothesis (H_1): That a CNN model can accurately and rapidly identify fluorescence in vivo confocal micrographs of high diagnostic quality.

Intra-oral confocal microscopy faces challenges in capturing high-quality in vivo images due to factors such as patient movement, saliva interference, and operator hand movement (Ramani et al., 2023; Yap et al., 2023). These issues, combined with sub-optimal contrast agent exposure, can lead to artifacts and distortions in confocal micrographs. To address this, a quality filtering step using convolutional neural networks (CNNs) was implemented before diagnostic analysis. CNNs, known for their effectiveness in image recognition tasks, can analyse image pixels to detect patterns and features (LeCun et al., 2015a).

The study employed transfer learning utilising the pre-trained Inception_V3 CNN model, which has demonstrated high performance on the ImageNet dataset (Deng, Dong, Socher, Li, Kai, et al., 2009; Huh et al., 2016). The CNN was developed using two approaches: one using MATLAB's Deep Network Designer with empirically chosen hyperparameters, and another using PyTorch with hyperparameter optimisation through a grid search algorithm and 5-fold cross-validation.

Despite consistent architectures and hyperparameters, performance differences were observed between MATLAB and PyTorch implementations. These variations can be attributed to differences in backend implementations, default settings, data handling, and preprocessing pipelines (Kim, 2017; Paszke et al., 2019). The approach of comparing CNNs developed in both environments depicted the impact of development environment on model performance.

The developed Quality Micrograph Refiner (QMR) CNNs showed high accuracy and F1 scores, with the PyTorch QMR outperforming the MATLAB version. The PyTorch QMR demonstrated high classification F1 scores across various oral sites, including

those with high OSCC incidence. However, it performed moderately on hard palate images, possibly due to a small sample size, access difficulties, and tissue keratinisation.

The in vivo confocal microscopy of the oral mucosa in this study yielded only 21.63% diagnostically useful images, with non-keratinised areas generally producing higher quality images than keratinised regions. This finding suggests a potential link between tissue keratinisation and image quality, highlighting the need for further research to refine quality filtering models and improve diagnostic imaging techniques in oral confocal microscopy.

Future research could focus on enhancing CNN models through advanced hyperparameter optimisation techniques, and architectural enhancements (Abdullah et al., 2022; He et al., 2016; M. Tan & Le, 2019). The quality filtering approach developed in this study has potential applications in standardising imaging techniques and improving operator training across various medical imaging fields.

An automated quality filtering approach involving CNNs when combined with real-time user feedback systems, could improve operator efficiency, assist in training new personnel, and facilitate quality control standardisation across diverse clinical and research settings. This quality filtering protocol has potential to be used in tandem with diagnostic triage processes to streamline real-time clinical diagnosis.

Since a CNN model could accurately and rapidly identify fluorescence in vivo confocal micrographs of high diagnostic quality the null hypothesis was rejected.

9.3. Hypothesis 3: Machine learning diagnostic analysis of human identified qualitative features

Aim: To develop machine learning models that can accurately identify oral potentially malignant disorders and oral squamous cell carcinoma based on human observed qualitative features observed on in vivo captured fluorescence confocal endomicroscopy images.

Null Hypothesis (H_0): That machine learning using human observed qualitative features of in vivo captured fluorescence confocal endomicroscopy images cannot accurately identify instances of oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That machine learning using human observed qualitative features of in vivo captured fluorescence confocal endomicroscopy images can accurately identify instances of oral potentially malignant disorders and oral squamous cell carcinoma.

This chapter describes a diagnostic study focusing on microscopic oral epithelial qualitative features as inputs for machine learning (ML) models to classify oral potentially malignant disorders in the form of lichenoid lesions and oral epithelial dysplasia (OED) along with oral squamous cell carcinoma (OSCC) (Ramani et al., 2023). The study used human-identified features such as cell size homogeneity, cell crowding, nucleus size homogeneity, nuclear crowding, and presence of fluorescence granules to train ML models for diagnostic classification.

The models showed inconsistent and unreliable performance in predicting lichenoid, OED, and OSCC cases, often miscategorising them as 'no dysplasia'. However, statistically significant associations were found between diagnostic categories and nuclear and cellular features, aligning with known microscopic markers of oral carcinogenesis (Odell et al., 2021b). These features, including altered cell proliferation, loss of epithelial organisation, and cytological atypia, correspond to observations in confocal micrographs and previous histopathological findings (Odell et al., 2021b).

Four different ML model types were used for robustness: logistic regression, support vector machines, random forest, and XGBoost. Despite covering a range of ML approaches, all models performed poorly across both binary and multi-category feature sets for both contrast agent datasets. The models showed high sensitivity for 'no

dysplasia' cases but low precision and specificity, with particularly poor performance for lichenoid cases.

The study's limitations included potential inconsistencies in human annotations, which can introduce noise in training data and impact ML model performance (Koçak et al., 2025; Yang et al., 2023). The novelty of fluorescence in vivo confocal microscopy and lack of specific standards for oral epithelium feature identification added to the annotation challenges. Additionally, the labour-intensive nature of human annotation led to a small dataset with class imbalance, potentially limiting ML model performance (Zhang & Qie, 2023).

Future improvements could involve using multiple expert annotators to reduce bias and exploring other ML models such as Bayesian approaches and instance-based neighbours classifiers (Abdullah et al., 2022). The results indicate that these ML models using human-identified qualitative features are not currently suitable for clinical use. Future work might focus on developing and validating a quantitative analysis methodology with machine-augmented systematic annotation of features, given the established link between oral epithelial features on fluorescence in vivo confocal microscopy and OED & OSCC.

Chairside clinical diagnostic systems that integrate human expertise with machine learning hold significant promise in improving early detection and triage of OED and OSCC. By combining the pattern recognition and consistency of machine learning with the contextual judgment and clinical experience of dental practitioners, such hybrid systems can enhance diagnostic accuracy, reduce variability, and support informed decision-making. This collaborative model also ensures accountability and patient safety while promoting the practical integration of advanced imaging tools, like confocal microscopy, into routine dental workflows. Additionally, the autonomy of clinicians in this AI powered solution can be valuable for improving patient trust.

Since the machine learning models developed in this study using human observed qualitative features of in vivo captured fluorescence confocal endomicroscopy images could not accurately identify instances of OPMDs and OSCC the null hypothesis was accepted.

9.4. Hypothesis 4: Machine learning diagnostic analysis of feature extraction segmented nuclei

Aim: To develop, evaluate, and compare ML models for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images.

Null Hypothesis (H_0): That ML models developed for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images cannot accurately detect oral potentially malignant disorders and oral squamous cell carcinoma.

Alternative Hypothesis (H_1): That ML models developed for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images can accurately detect oral potentially malignant disorders and oral squamous cell carcinoma

This study aimed to develop and validate a quantitative analysis method using machine learning for assessing oral potentially malignant disorders in the form of lichenoid lesions and oral epithelial dysplasia (OED) along with oral squamous cell carcinoma (OSCC), following poor results from previous human-annotated analyses. The methodology centred on comparing U-Net inspired segmentation algorithms, with StarDist 2D significantly outperforming Cellpose 2D ($p>0.015$) (Stringer et al., 2021; Weigert et al., 2020). The StarDist 2D model showed high median MMS (0.85) but poor median MTS (0.31) in nucleus segmentation, using metrics such as area, pixel intensity, integrated density, circularity, and aspect ratio for nucleus measurement.

Three quantitative analysis approaches were employed: mean values of nuclear measurements, k-means clustering of nuclei measurements, and distances between nuclei. Four types of machine learning models, including logistic regression, support vector machine, random forest, and XGBoost, were tested on these approaches. The acriflavine dataset (91,550 nuclei, 1322 images) revealed that high-risk nuclei were the brightest with the highest median mean grayness and integrated density, while low-risk nuclei showed the greatest variability in brightness. The fluorescein dataset (10,722 nuclei, 640 images) indicated that high-risk nuclei were larger, brighter, and had more variable brightness, while lichenoid nuclei exhibited the lowest fluorescence brightness overall.

The best overall performance came from the fluorescein random forest model using the approach involving mean values of nuclear measurements, achieving moderate to high AUC values (0.72 to 0.87) but only correctly identifying about half of the OED and OSCC cases. This performance, while promising, fell short of clinical requirements for predicting biopsy needs for OED and OSCC.

Several important factors influenced the study's outcomes. The choice of contrast agent significantly impacted feature extraction, with acriflavine binding directly to nucleic acids while fluorescein diffused passively into cells (Piorecka et al., 2022; Robertson et al., 2013). This difference likely contributed to variations in pixel brightness values between the two datasets. Lighting conditions and field of view selection could also contribute as crucial factors affecting feature extraction consistency and reliability. The study primarily focused on nuclear features, potentially overlooking other important cellular and architectural changes visible in histopathology.

The limited efficacy of topically applied contrast agents was a challenge, with some confocal micrographs showing an absence of nuclei. This limitation suggests the potential benefit of more targeted biomarker contrast agents, such as Poly ADP-ribose Polymerase inhibitor-Fluorescent (PARPi-Fl), which has shown promise in imaging advanced oral cancer cases (Demetrio de Souza Franca et al., 2021).

Despite not achieving the intended clinical applicability, the study's quantitative measurements identified patterns in nuclear characteristics that might extend to cellular and architectural changes. These findings align with previous literature highlighting increased nuclear cell density in cancer cells (Uttam et al., 2015) and the WHO grading system for dysplastic nuclei in histopathology and immunohistochemistry (Odell et al., 2021a).

The study highlights the complexities involved in translating microscopic image analysis into clinically relevant diagnostic tools. It suggests that further research using convolutional neural networks (CNN) might be beneficial to consider all features present in the fluorescence confocal micrographs, potentially improving the accuracy and applicability of the method in clinical settings. Additionally, the study underscores the need for more advanced contrast agents and imaging techniques to better visualise and quantify the full range of cellular and architectural changes associated with oral dysplasia and cancer.

The quantitative feature extraction approach of employing machine-identified features can introduce objectivity and reproducible into clinical diagnosis workflows. The explainable dimension of this feature extraction approach is easier for clinicians and imaging operators to interpret as opposed to the ‘black box’ CNN approach. Improving this methodology for confocal microscopy image analysis could potentially improve our understanding of the morphological changes seen in microscopic components of the oral epithelium in various disease processes.

The null hypothesis for this study was accepted since the ML models developed for diagnostic classification of feature extracted data of epithelial cell nuclei measurements from fluorescence human in vivo captured confocal endomicroscopy images could not accurately detect lichenoid lesions, OED and OSCC.

9.5. Hypothesis 5: Deep learning classification diagnostic triage models

Aim: To develop and test deep learning convolutional neural network (CNN) models with on fluorescence in vivo confocal microscopy images of the oral mucosa for the accurate detection of oral potentially malignant disorders and oral squamous cell carcinoma.

Null Hypothesis (H_0): That deep learning CNN models cannot accurately detect oral potentially malignant disorders and oral squamous cell carcinoma in fluorescence in vivo confocal microscopy images of the oral mucosa.

Alternative Hypothesis (H_1): That deep learning CNN models can accurately detect oral potentially malignant disorders and oral squamous cell carcinoma in fluorescence in vivo confocal microscopy images of the oral mucosa.

This chapter aimed to develop diagnostic triage models for the human fluorescence in vivo confocal microscopy images. This study highlights the variation in performance of CNN across different contrast agents (acriflavine & fluorescein) and development platforms (MATLAB & PyTorch) for detection of oral potentially malignant disorders in the form of lichenoid lesions and oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) on a fluorescence in vivo confocal microscopy dataset collected of the human oral mucosa.

The fluorescein CNNs showed consistent performance between MATLAB and PyTorch implementations, with the PyTorch version demonstrating balanced performance across diagnostic categories. In contrast, there was a significant difference in test performance between the acriflavine models in MATLAB and PyTorch implementations, emphasising the impact of implementation details on model performance (LeCun et al., 2015a).

The acriflavine MATLAB diagnostic triage CNN model performed best overall, excelling at detecting high-grade OED & OSCC with 100% sensitivity and a low false positive rate (26.1%). However, the perfect AUC of 1 for high-risk images raises concerns about potential overfitting, particularly given the small sample size for this class.

To improve clinical applicability, future studies should use larger, more balanced datasets across all diagnostic categories. The transformative impact of large-scale

datasets on CNN performance was demonstrated by the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Deng, Dong, Socher, Li, Kai, et al., 2009; Szegedy et al., 2016). A prospective multi-centre trial could enhance the generalisability of the models developed.

The black box nature of CNNs highlights potential issues with clinical implementation, including automation complacency and confirmation bias among clinicians (Challen et al., 2019; Parasuraman & Manzey, 2010). These concerns emphasise the need for AI literacy among clinicians to preserve critical thinking in clinical decision-making.

Despite its rapid classification speed, this technology is not intended to replace standard biopsy-histopathology procedures at this stage. Instead, it offers the potential for high-rate, non-invasive sampling across multiple lesions, potentially improving diagnostic precision and patient monitoring efficiency. However, further validation through randomised controlled trials is necessary before considering software integration and clinical application (Davis et al., 2017).

Due to the rapid and precise prediction properties, CNNs are a natural fit for an automated chairside diagnostic system. It's direct and intuitive output would be straightforward for clinicians and imaging operators to incorporate into their diagnostic workflows. The 'black box' nature of these models does make it challenging to interpret their predictions and more work is needed in this area to improve trust. As with any other intervention, the CNN diagnostic approach needs to be validated using a randomised controlled trial in a large multi-centre dataset to assess its generalisability.

Since the CNN models developed in this study could accurately and rapidly detect OED and OSCC in fluorescence in vivo confocal microscopy images of the oral mucosa the null hypothesis was rejected.

9.6. Hypothesis 6: Diagnostic triage deep learning models in a murine model of oral carcinogenesis

Aim: To develop and evaluate the performance of deep learning convolutional neural network (CNN) models for the accurate classification of oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

Null Hypothesis (H_0): That a trained deep learning model cannot accurately detect OED and OSCC using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

Alternative Hypothesis (H_1): That a trained deep learning model can accurately detect OED and OSCC using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model.

The research aimed to test the CNN diagnostic triage approach previously used in human confocal microscopy datasets on a murine model, recognising the potential benefits of non-invasive AI diagnostics in animal research. CNNs were developed using MATLAB and PyTorch deep learning frameworks. Both models misclassified majority of the OED & OSCC cases and were therefore considered unsuitable for clinical application. This performance was unsatisfactory and was lower than that achieved with the human image dataset.

Several challenges were identified during the study. Image variability was a significant issue, with murine confocal micrographs showing considerable variation in appearance, ranging from superficial to deep cross-sections of the tissue (Liu & Zhang, 2019). In vivo imaging difficulties arose due to the small size of the mouse oral cavity, which posed challenges for the 3.5mm diameter probe of the ViewnVivo system (Optiscan Imaging, Australia) (Celentano, Rickard, Low, Silke, Mohammed, Moslemi, Ramani, Franca, Reiner, McCullough, et al., 2025). Low inter-class variation was observed, as distinct tissue types among diagnostic categories appeared visually similar in fluorescence confocal micrographs of the 4NQO oral carcinogenesis model. Additionally, the 'black box' nature of CNNs with millions of parameters prevented interpretation of reasoning and feature selection (LeCun et al., 2015a).

To improve the approach, there are several strategies. These include collecting a larger, more balanced sample size of images from varied mice, trialling other CNN

architectures to explore potential identification of new features and applying advances in explainable AI to fine-tune CNN performance.

Despite suboptimal initial results, the integration of in vivo confocal microscopy with CNN models presents a promising strategy for oral cancer detection in murine models. Confocal microscopy provides real-time, high-resolution "optical biopsies" of the mouse oral cavity, allowing for repeated imaging of the same lesions over time (Rangrez, Bussau, Ifrit, & Delaney, 2021). This non-invasive approach is particularly valuable for longitudinal studies tracking disease progression or therapeutic responses within individual animals, potentially reducing the need for large animal cohorts. Furthermore, CNNs can be iteratively refined and improved with more data and advanced architectures, meaning that even moderately poor initial results serve as valuable benchmarks for future development (LeCun et al., 2015a).

With non-invasive imaging, we can monitor the same group of animals over time by taking measurements before and after within the same individual. This way, each animal serves as its own control. By comparing changes within the same animal, rather than across different ones, we avoid the biological variability that can impact results. The outcome is a clearer, more accurate picture of how a disease progresses or how a treatment works. This approach is also more efficient and ethically responsible. It reduces the number of animals needed, cuts down on costs and resources, and importantly, supports the 3Rs in animal research: Replace, Reduce, and Refine as defined by the National Health and Medical Research Council (NHMRC) of Australia (NHMRC, 2019). Because the imaging is non-invasive, animals are spared repeated invasive procedures and don’t need to be euthanised at each stage, making the entire process less stressful and more humane.

Since the trained deep learning model could not accurately detect OED and oral cancer using fluorescence in vivo confocal microscopy imaging of an oral cancer mouse model the null hypothesis was accepted.

9.7. Limitations

While all the studies in this dissertation make valuable insights in the field of image analysis in oral oncology, they have several notable limitations which should be considered when interpreting their results.

Despite having 59 patients to draw from for the human confocal microscopy dataset, the number of images used to develop and validate all AI models was potentially a limiting factor. There is an ongoing debate in the AI scientific community regarding the appropriate sample size for developing the most efficient CNN models. The research team that developed the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition and the industry and academic research teams that participated in the challenge highlighted the impact of the 14 million image dataset on high CNN performance (Deng, Dong, Socher, Li, Kai, et al., 2009). While other academics have offered a rigorous theoretical foundations for the empirical observation that CNNs often require fewer samples to train effectively compared to other ML models (Du et al., 2018).

Assuming a large dataset is a core requirement of specialised CNNs and considering the ImageNet models were developed using millions of images for specialist tasks, our dataset of close to two thousand images appears to be lacking. This is in part due to over seven thousand images being needed to be discarded out of our original dataset of over nine thousand images for poor quality. In vivo imaging is challenging with a mobile patient, a high requirement for hand stability & operator expertise, and patient saliva potentially interfering with the imaging focus. This lack of a large image volume was somewhat balanced by transfer learning where an Inception_v3 model which was pre-trained on 14 million images from its participation in the ILSVRC was allowed to retain its previous learnings and pivoted to adapt to our specialised dataset instead of starting from scratch (Szegedy et al., 2016). The learning curve for in vivo capture using a probe-based fluorescence confocal microscope is high and standardised imaging protocols need to be designed and validation for consistent imaging as reported by the systematic review studies in the Chapter 2.

This collection of data from only one centre for human subjects (RDHM, Melbourne) limits the generalisability of the results. AI models especially being developed in different fields of dentistry have noted the biases that can contaminate the learning process (Park, Schwendicke, Krois, Huh, & Lee, 2023). Producing a model that learns to apply its predictions to the entire target population of patients requires a comprehensive representation of all possible demographics that constitute said population (Ma et al., 2022). This not only includes patients of different genders, cultures, and ages but also socio-economic status. While representing the entire oral cancer population might not have been feasible within the scope of the studies in the

current dissertation, it is a limitation. The results must be interpretated with the caveat that this represents a sample of the population of patients with oral mucosal conditions living in Victoria, Australia. They cannot be directly extrapolated and applied to all global population samples without further validation.

In addition to the relatively modest sample size, the distribution of images across the diagnostic categories were imbalanced with the dataset comprising of 37.46% non-dysplastic tissue images, 26.77% lichenoid images, 25.87% low risk images, and 9.88% high risk images. Since participants for this study were recruited based on their attendance at the Oral Medicine clinic at the Royal Dental Hospital Melbourne (Melbourne, Australia) the rates of OSCC and high-grade OED were low. The test sets across all diagnostic studies were selected randomly but in proportion to the underlying class distribution. While this approach preserves ecological validity and prevents artificial inflation of rare categories, it also resulted in relatively few test cases for certain diagnostic groups. As a consequence, performance estimates for under-represented classes, such as high-grade OED and OSCC, are less stable and may be disproportionately affected by individual misclassifications (Cancer, 2023). This limitation reflects a common trade-off between maintaining realistic class prevalence and ensuring sufficient sample size for robust per-class evaluation.

Datasets with an imbalanced number of instances in each classification category impacts the performance of ML models. Algorithms that look at fewer training examples from a specific category have a lower probability of learning important features for that class compared to other classes (Thabtah, Hammoud, Kamalov, & Gonsalves, 2020). Additionally, the performance validation of ML models trained on imbalanced datasets needs further consideration for the specific metric being used. Certain metrics such as accuracy are not sensitive to changes in classification predictions in minority classes (Reinke et al., 2024). This could lead to an overestimation of model performance if the model makes correct predictions only for the majority classes. To improve interpretation of the results in our study the class imbalance was acknowledged and a range of metrics including sensitivity, specificity, precision, F1 score and AUROC were reported.

The field of microscopy image analysis is rapidly growing with a wealth of novel tools being available and accessible for research. The open-source segmentation models tested in this study for extracting and measuring epithelial cell nuclei were Cellpose 2D and StarDist 2D. However, there are other publicly available image analysis tools for this specialised task such as CellProfiler, DeepCell, and InstanSeg (Bannon et al., 2021; Goldsborough et al., 2024; Stirling et al., 2021). Similarly with regards to the CNN experiments, while Inception_V3 was developed on our dataset, it is far from the only open-source CNN model available. CNN architectures such as ResNet, EfficientNet, DenseNet, and so on have differences in the way the feature maps are mathematically

computed (He et al., 2016; G. Huang, Liu, Van Der Maaten, & Weinberger, 2017; M. Tan & Le, 2019).

Additionally, the novel transformer architecture AI models popularly used in large language models (LLM) such as ChatGPT, DeepSeek, and Google Gemini among others has shown great potential in image analysis (Johnson, Alyasiri, Akhtom, & Johnson, 2023; Lu et al., 2024; Saab et al., 2024). This extends to their high performance in dental imaging (Büttner, Leser, Schneider, & Schwendicke, 2024). While these models have performed well within their respective test conditions due to the subject specialised nature of the dataset analysed in this dissertation, there is no guarantee they would perform well under the same conditions (Johnson et al., 2023; Lu et al., 2024; Saab et al., 2024). However, a constantly growing list of new AI models were not included in this dissertation due to time and computational costs. A vital consideration is environmental burden of running training for several AI models for prolonged durations using high performance computing (Rillig, Ågerstrand, Bi, Gould, & Sauerland, 2023). Training advanced AI models, such as transformers, demands significant computational resources, leading to high energy usage. For instance, training a modern LLM named GPT-3 required approximately 1,287 megawatt-hours of electricity, resulting in 502,000 kg of CO_2 emissions (Iftikhar & Davy, 2024). This level of energy consumption is comparable to the lifetime emissions of multiple petroleum-powered vehicles (Strubell, Ganesh, & McCallum, 2020).

CNNs are often termed as ‘black boxes’ due to the inherent complexity of the mathematical functions employed by these models for extracting features for image identification. This prevents us from simply pulling the trained model apart to understand the reasoning for classification decisions. This scientific pursuit for understanding how machine learning mechanisms work has been given the new name of explainable artificial intelligence (XAI) by researchers at the Defense Advanced Research Projects Agency (DARPA) as an agency of the United States government (Gunning & Aha, 2019). This term has caught on with the rest of the AI community and novel strategies and tools are being developed for XAI across different model types and architectures (Holzinger, Saranti, Molnar, Biecek, & Samek, 2020).

No XAI methods were tested in this dissertation to attempt to understand the decision reasoning of the ML models developed. This is in part due to the recency of these methods being introduced into interpreting CNNs in microscopy image analysis. The specific dataset of fluorescence in vivo confocal microscopy images analysed in this dissertation have no formally validated and rigorously tested criteria for identification of specific features and landmarks with topical acriflavine and fluorescein as contrast agents. This lack of a strong understanding of the features themselves mitigates the potential benefits of employing an XAI method at this stage. Standards are being discussed and developed for the appropriate use of XAI in dentistry with the vision of

employing them on all developed models to improve understanding and validation (Ma et al., 2022).

When it comes to integrating these AI systems into patient diagnostic workflows clinicians interacting with AI systems are vulnerable to automation complacency and confirmation bias, particularly when systems appear highly reliable (Davis et al., 2017; Parasuraman & Manzey, 2010). These cognitive pitfalls can lead to over-trust, reduced vigilance, and failure to detect AI errors, especially under high workload conditions. Addressing these risks requires AI systems to communicate uncertainty effectively and support active human oversight (Challen et al., 2019).

9.8. Future directions

The next step in advancement from the studies described in this dissertation could focus on addressing their limitations. A prospective trial could be designed where a larger fluorescence in vivo confocal microscopy dataset is collected with the specific aim of AI analysis. This would involve efforts to collect a balanced dataset across different OED and OSCC samples. Additionally, representing a variety of demographics by conducting a multi-centre trial would enhance the generalisability of the models developed.

As image recognition technology advances CNNs are currently being challenged by a new type of AI model architecture called Vision Transformers (ViT) which is a subset of the transformed architecture used in state of the art large language models (Vaswani et al., 2017). CNNs have a strong inductive bias for local patterns. Their convolutional layers are designed to find features like edges, textures, and corners within small, localized regions in the image. This works well and is efficient, but it means the network must build a global understanding of the image hierarchically, layer by layer. The core idea of the ViT is to completely abandon this convolutional inductive bias. Instead, it treats an image as a sequence of flattened patches, much like a sentence is a sequence of words (Dosovitskiy et al., 2020). A transformer encoder then uses a self-attention mechanism to determine the relationship between every patch and every other patch, regardless of their distance (Vaswani et al., 2017).

Shifting from CNNs to ViTs opens up exciting new possibilities, especially because ViTs can capture global patterns in images through self-attention and understand context within images (Dosovitskiy et al., 2020). This gives them an edge over traditional CNNs particularly when working with very large datasets. However, the trade-off is that ViTs are more computationally demanding and need a lot of data to make up for their lack of built-in inductive biases, like the spatial hierarchies that CNNs naturally model (Dosovitskiy et al., 2020). That means for smaller datasets such as the one in the current study or resource-limited settings, CNNs still hold their ground as a highly efficient and effective choice. In future research, hybrid models that blend CNNs' efficiency with the global perspective of ViTs could be explored for strong performance without demanding huge amounts of data or compute.

Expanding the scope of AI research in medical domain is characterised by risks relating to data privacy regulations (such as the Privacy act 1988) which often restrict the sharing of patient information across institutions. Some of this risk was mitigated in the studies in the current dissertation by de-identification and a thorough ethics approval process. A more systematic method for protecting data privacy is Federated learning (L. Li, Fan, Tse, & Lin, 2020). This form of AI development offers a solution by allowing ML models to be trained across multiple healthcare sites without moving raw data off-

site. Instead, each hospital or device keeps data locally and only exchanges model parameters or updates, thus greatly reducing privacy risks (L. Li et al., 2020). This approach of distributing models between research institutions is being explored for AI image analysis in dentistry (Rischke et al., 2022). A recent study by Schneider et al. (2023), compared federated learning to local learning (training models on isolated data from each centre) and central learning (training on centrally pooled data) for automated tooth segmentation on panoramic radiographs (Schneider et al., 2023). The authors found that if the data can be pooled with data sharing agreements, then central learning outperforms federated and local learning. However, in all other cases federated learning outperformed other approaches (Schneider et al., 2023). The Australian Research Data Commons (ARDC) initiated a pathfinder project to explore the potential of Federated Learning (FL) for sensitive health-related data in Australia (Holloway et al., 2024). They recommended support for community and governance frameworks for federated learning adoption in Australia to promote data equity, security, and ethical use across research (Holloway et al., 2024). This highlights the potential for federated learning in combining data from healthcare institutions all across the world to generate optimum generalisable AI models.

A solution for enhancing smaller datasets while reducing the risk of data privacy breaches and avoiding complex data sharing agreements is synthetic data augmentation. Synthetic data refers to artificially generated information that mimics real patient data but does not correspond to actual individuals (Nikolenko, 2021). In medical ML, synthetic data is used to supplement or replace real clinical data for training, testing, and validating algorithms (Gonzales, Guruswamy, & Smith, 2023). The concept of synthetic data addresses the current limitations of dental data and recommendations by international dental regulatory bodies by providing privacy-preserving yet representative cases for AI research and development (Feher, Tussie, & Giannobile, 2024). While synthetic data might be a promising solution, it is a relatively new approach that requires further research to validate its reliability and the impact on AI models. The research community must establish robust evaluation frameworks and conduct comprehensive validation studies to ensure these models perform safely and effectively in real-world clinical contexts.

In terms of scaling up the types of AI models used, a robust approach to progression would involve developing and validating a wider variety of image analysis segmentation tools such as CellProfiler, DeepCell, and InstanSeg (Bannon et al., 2021; Goldsborough et al., 2024; Stirling et al., 2021). Additionally, a range of CNN models can be trialled such as ResNet, EfficientNet, DenseNet, and other upcoming algorithms (He et al., 2016; G. Huang et al., 2017; M. Tan & Le, 2019). While controlling for the computational and environmental carbon footprint considerations even the transformer model architecture can be developed on a larger dataset. The defining characteristic of this architecture that sets it apart from CNNs is the attention mechanism (Vaswani et al., 2017). The attention mechanism is a method that helps a model decide which parts of

the input information are most important for producing each part of the output. In practice, the mechanism builds three sets of numbers from the input: one set is used to look for relevant information, one set helps decide how important other parts are, and the third set holds the information to be combined (Vaswani et al., 2017). This technology that powers popular LLMs such as ChatGPT and DeepSeek translates well to image recognition tasks as understanding the context of neighbouring groups of pixels help the model understand landmarks on the image. Transformer architecture models have potential in use for confocal micrographs of the oral epithelium if they can identify signs of dysplasia.

The complexity of ML models is often attributed to them overfitting to the training data. This involves the model becoming overly familiar with the variations found in the sample data, especially the random noise and learning aspects of the training data that are redundant and not relevant to the task at hand. Such learning is expected to be detrimental when the model is tested on an unseen dataset with different variations random noise compared to what the model had memorised from the training data. This is especially true with modern day CNN models with millions of parameters. For instance, the Inception_V3 model selected in the studies in this dissertation has approximately 23.8 million parameters (Szegedy et al., 2016). This number vastly overshadows the 1983 training images across both contrast agents. Such ML models that have significantly more parameters than training datapoints are called over-parameterised and could either lead to overfitting or could generalise well depending on the learning task (Allen-Zhu, Li, & Liang, 2019). However, in the past few years this fundamental theory has been thoroughly challenged. We have entered a terra incognita or ‘unknown land’ in AI development, where the most complex models which find the most convoluted relationships within the limited data points made available to them. This helps them go beyond memorising the dataset (overfitting) and into the territory of developing a new understanding which the top AI scientists are striving to understand (Ananthaswamy, 2024). This phenomenon goes against the very fibre of fundamental machine learning theory, and we have entered the wild west of experimental AI research with limited assumptions using past knowledge.

9.9. Conclusions

The analysis of fluorescence in vivo confocal microscopy images with machine learning via convolutional neural networks can provide a rapid and precise diagnosis of oral potentially malignant disorders and oral squamous cell carcinoma in real-time.

A systematic review of confocal microscopy in oral cancer diagnosis demonstrated the utility of digital biopsies for tissue level diagnosis of OSCC and dysplasia in both in vivo and ex vivo specimens from the oral cavity. There was heterogeneity across the studies regarding the assessment of oral mucosa tissue. Both qualitative and quantitative confocal assessment methodologies have been explored, the latter highlighting the potential of future machine-augmented diagnostic precision.

Within the scope of this work CNNs specially developed for quality filtering and diagnostic triage on specialist fluorescence in vivo captured confocal microscopy images showed a high rate of correct identification. This approach involves a tandem CNN workflow which is characterised by initial quality filtering followed by diagnostic triage.

The developed qualitative human-identified features of epithelial components and quantitative feature extraction analysis of nuclear measurements identified some statistically significant relationships with OED and OSCC. The quantitative feature extraction approach involving measuring epithelial cell nuclei also identified statistically significant differences between nuclei in low-grade and high-grade OED and OSCC images compared to other diagnostic groups. High-grade OED and OSCC nuclei appeared to be brighter and larger than other nuclei with an increased variability. However, the ML models developed on both of these approaches failed to accurately predict the associated condition and were ultimately deemed unfit for clinical diagnostic triage of OED and OSCC.

While the CNNs in this murine confocal imaging study performed poorly at triaging mice for potentially malignant and malignant oral lesions, these results are promising signs for using this technology in preventing animal sacrifice in novel cancer detection and therapy testing studies. The combination of in vivo confocal microscopy and AI-based image analysis holds significant potential for advancing oral cancer research while adhering to more ethical research standards.

While the CNNs developed in this dissertation showed immense potential in being incorporated into digital dentistry workflows for cancer screening, surgical biopsies with histopathology are still the current standard of care for the diagnosis of OED and OSCC. The technology introduced in this work could dramatically increase the sampling done by clinicians of oral mucosal conditions in a non-invasive manner within

a few seconds. This could help triage a range of oral mucosal conditions and monitor them a regular interval without introducing any biopsy wounds.

The complex landscape of performance variations across microscopy imaging hardware, staining techniques, types of machine learning approaches, algorithm development frameworks and diagnostic categories calls for further investigation. Future work should focus on understanding and mitigating the sources of these variations to develop more robust, reliable, and clinically applicable AI-driven diagnostic tools for oral cancer detection.

The analysis of fluorescence in vivo confocal microscopy images using CNNs demonstrates considerable promise for rapid, non-invasive diagnosis of OPMDs and OSCC, yet challenges remain in achieving clinically reliable accuracy. While a systematic review of literature affirmed the diagnostic potential of digital biopsies, the heterogeneity in imaging methods and analytical approaches continues to limit comparability and reproducibility. Qualitative and quantitative analyses of the image data revealed significant nuclear and epithelial differences across OED and OSCC grades, underscoring measurable diagnostic signals. However, these insights did not translate to accurate risk assessment predictions by machine learning predictive algorithms. This study's tandem CNN workflow for image quality filtering and diagnostic triage performed well in image recognition but struggled with accurate disease classification. This reflects the biological and technical variability inherent in such data. Although not yet ready for clinical deployment, integrating AI with fluorescence in vivo confocal microscopy offers a transformative, ethical route towards real-time oral cancer detection and monitoring, meriting further refinement and standardisation.

The scope for this line of research is growing endlessly with no boundaries in sight. The only limits are set by our imagination and hopefully human ethics!

10. REFERENCES

Abdullah, A. A., Hassan, M. M., & Mustafa, Y. T. (2022). A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access, 10*, 36538-36562.

Aghbari, S. M. H., Abushouk, A. I., Attia, A., Elmaraezy, A., Menshawy, A., Ahmed, M. S., . . . Ahmed, E. M. (2017). Malignant transformation of oral lichen planus and oral lichenoid lesions: A meta-analysis of 20095 patient data. *Oral Oncology, 68*, 92-102.

AIHW. (2021). *Cancer data in Australia* *https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/cancer-summary-data-visualisation* Retrieved from

Alessi, S. S., Nico, M. M. S., Fernandes, J. D., & Lourenço, S. V. (2013). Reflectance confocal microscopy as a new tool in thein vivoevaluation of desquamative gingivitis: patterns in mucous membrane pemphigoid, pemphigus vulgaris and oral lichen planus. *British Journal of Dermatology, 168*(2), 257-264. doi:10.1111/bjd.12021

Alibrahim, H., & Ludwig, S. A. (2021). *Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization.* Paper presented at the 2021 IEEE Congress on Evolutionary Computation (CEC).

Allen-Zhu, Z., Li, Y., & Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in Neural Information Processing Systems, 32*.

Ananthaswamy, A. (2024). *Why machines learn: the elegant maths behind modern AI*: Random House.

Anderson, A. W., Marinovich, M. L., Houssami, N., Lowry, K. P., Elmore, J. G., Buist, D. S., . . . Lee, C. I. (2022). Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: a systematic review. *Journal of the American College of Radiology, 19*(2), 259-273.

Anuthama, K., Sherlin, H. J., Anuja, N., Ramani, P., Premkumar, P., & Chandrasekar, T. (2010). Characterization of different tissue changes in normal, betel chewers, potentially malignant lesions, conditions and oral squamous cell carcinoma using reflectance confocal microscopy: Correlation with routine histopathology. *Oral oncology, 46*(4), 232-248.

Arsiwala-Scheppach, L. T., Chaurasia, A., Mueller, A., Krois, J., & Schwendicke, F. (2023). Machine learning in dentistry: a scoping review. *Journal of Clinical Medicine, 12*(3), 937.

Banerjee, A., Kamath, V. V., Lavanya, R., Shruthi, S., & Deepa, M. (2015). Mathematical (diagnostic) algorithms in the digitization of oral histopathology: The new frontier in histopathological diagnosis. *Journal of Dental Research Reviews, 2*(2), 97-101.

Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., . . . Osterman, E. (2021). DeepCell Kiosk: scaling deep learning–enabled cellular image analysis with Kubernetes. *Nature methods, 18*(1), 43-45.

Beard, P. (2011). Biomedical photoacoustic imaging. *Interface focus, 1*(4), 602-631.

Belaldavar, C., Angadi, P. V., & Mudenagudi, U. (2024). QuPath for automated analysis of digital images of oral epithelial dysplasia. *Journal of Oral Maxillofacial Pathology, 28*(3), 381-386.

Bellando-Randone, S., Russo, E., Venerito, V., Matucci-Cerinic, M., Iannone, F., Tangaro, S., & Amedei, A. (2021). Exploring the oral microbiome in rheumatic diseases, state of art and future prospective in personalized medicine with an AI approach. *Journal of Personalized Medicine, 11*(7), 625.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research, 13*(2).

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4): Springer.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . De Vet, H. C. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277*(3), 826-832.

Bray, M.-A., Fraser, A. N., Hasaka, T. P., & Carpenter, A. E. (2012). Workflow and metrics for image quality control in large-scale high-content screens. *Journal of biomolecular screening, 17*(2), 266-274.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Büttner, M., Leser, U., Schneider, L., & Schwendicke, F. (2024). Natural language processing: chances and challenges in dentistry. *Journal of dentistry, 141*, 104796.

Büttner, M., Rokhshad, R., Brinz, J., Issa, J., Chaurasia, A., Uribe, S. E., . . . Schwendicke, F. (2024). Core outcomes measures in dental computer vision studies (DentalCOMS). *Journal of Dentistry, 150*, 105318.

Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., . . . McQuin, C. (2019). Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature methods, 16*(12), 1247-1253.

Camalan, S., Mahmood, H., Binol, H., Araujo, A. L. D., Santos-Silva, A. R., Vargas, P. A., . . . Gurcan, M. N. (2021). Convolutional neural network-based clinical predictors of oral dysplasia: Class activation map analysis of deep learning results. *Cancers, 13*(6), 1291.

Cancer, I. A. f. R. o. (2023). *IARC handbooks of cancer prevention* (Vol. 19): The Agency.

Canto, M. I., Anandasabapathy, S., Brugge, W., Falk, G. W., Dunbar, K. B., Zhang, Z., . . . Goldblum, J. (2014). In vivo endomicroscopy improves detection of Barrett's esophagus–related neoplasia: a multicenter international randomized controlled trial (with video). *Gastrointestinal endoscopy, 79*(2), 211-221.

Carlson, A. L., Gillenwater, A. M., Williams, M. D., El-Naggar, A. K., & Richards-Kortum, R. R. (2007a). Confocal microscopy and molecular-specific optical contrast agents for the detection of oral neoplasia. *Technology in cancer research & treatment, 6*(5), 361-374.

Carlson, A. L., Gillenwater, A. M., Williams, M. D., El-Naggar, A. K., & Richards-Kortum, R. R. (2007b). Confocal microscopy and molecular-specific optical contrast agents for the detection of oral neoplasia. *Technology in cancer research treatment, 6*(5), 361-374.

Carrozzo, M., Porter, S., Mercadante, V., & Fedele, S. (2019). Oral lichen planus: A disease or a spectrum of tissue reactions? Types, causes, diagnostic algorhythms, prognosis, management strategies. *Periodontology, 80*(1), 105-125.

Celentano, A., & Cirillo, N. (2024). Diseases with oral malignant potential: Need for change to inform research, policy, and practice. *Journal of Oral Pathology Medicine, 53*(8), 495-501.

Celentano, A., Rickard, J. A., Low, J., Silke, N., Mohammed, A. I., Moslemi, E., . . . McCullough, M. J. (2025). Enabling high-resolution diagnostic oral confocal laser endomicroscopy in mice. *Methods*.

Celentano, A., Rickard, J. A., Low, J., Silke, N., Mohammed, A. I., Moslemi, E., . . . Yap, T. (2025). Enabling high-resolution diagnostic oral confocal laser endomicroscopy in mice. *Methods*.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ quality safety, 28*(3), 231-237.

Chaturvedi, A. K., Udaltsova, N., Engels, E. A., Katzel, J. A., Yanik, E. L., Katki, H. A., . . . Silverberg, M. J. (2020). Oral leukoplakia and risk of progression to oral cancer: a population-based cohort study. *JNCI: Journal of the National Cancer Institute, 112*(10), 1047-1054.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chu, P. (2010). *Advances in solid state circuit technologies*. Vukovar, Croatia: Intech.

Clark, A., Collier, T., Lacy, A., Follen, M., Malpica, A., Gillenwater, A., & Richards-Kortum, R. (2002). Detection of dysplasia with near real time confocal microscopy. *Biomedical Sciences Instrumentation, 38*, 393-398.

Clark, A. L., Collier, A. M., Alizadeh-Naderi, T. G., Reza, El-Naggar, A. K., & Richards-Kortum, R. R. (2003). Confocal microscopy for real-time detection of oral cavity neoplasia. *Clinical Cancer Research, 9*(13), 4714-4721.

Contaldo, M., Di Stasio, D., Petruzzi, M., Serpico, R., & Lucchese, A. (2019). In vivo reflectance confocal microscopy of oral lichen planus. *International journal of dermatology, 58*(8), 940-945.

Contaldo, M., Lauritano, D., Carinci, F., Romano, A., Di Stasio, D., Lajolo, C., . . . Lucchese, A. (2020). Intraoral confocal microscopy of suspicious oral lesions: a prospective case series. *International journal of dermatology, 59*(1), 82-90.

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association, 24*(6), 1052-1061.

Delaney, P. M., Harris, M. R., & King, R. G. (1994). Fiber-optic laser scanning confocal microscope suitable for fluorescence imaging. *Applied Optics, 33*(4), 573-577.

Demetrio de Souza Franca, P., Kossatz, S., Brand, C., Karassawa Zanoni, D., Roberts, S., Guru, N., . . . Weber, W. A. (2021). A phase I study of a PARP1-targeted topical fluorophore for the detection of oral cancer. *European journal of nuclear medicine molecular imaging, 48*(11), 3618-3630.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai, L., & Li, F.-F. (2009, 2009). *ImageNet: A large-scale hierarchical image database*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database.* Paper presented at the 2009 IEEE conference on computer vision and pattern recognition.

Devi, T. G., & Patil, N. (2020). *Analysis & evaluation of Image filtering Noise reduction technique for Microscopic Images.* Paper presented at the 2020 International Conference on Innovative Trends in Information Technology (ICITIIT).

Dittberner, A., Rodner, E., Ortmann, W., Stadler, J., Schmidt, C., Petersen, I., . . . Guntinas–Lichius, O. (2016a). Automated analysis of confocal laser endomicroscopy images to detect head and neck cancer. *Head Neck, 38*(S1), E1419-E1426.

Dittberner, A., Rodner, E., Ortmann, W., Stadler, J., Schmidt, C., Petersen, I., . . . Guntinas–Lichius, O. (2016b). Automated analysis of confocal laser endomicroscopy images to detect head and neck cancer. *Head & Neck, 38*(S1), E1419-E1426.

Dittberner, A., Ziadat, R., Hoffmann, F., Pertzborn, D., Guntinas-Lichius, O., & Gassler, N. (2021). Fluorescein-Guided Panendoscopy for Head and Neck Cancer Using Handheld Probe-Based Confocal Laser Endomicroscopy: A Pilot Study. *Frontiers in Oncology, 11*, 671880. doi:http://dx.doi.org/10.3389/fonc.2021.671880

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv, 2*(11929).

Dost, F., Lê Cao, K., Ford, P., Ades, C., & Farah, C. (2014). Malignant transformation of oral epithelial dysplasia: a real-world evaluation of histopathologic grading. *Oral surgery, oral medicine, oral pathology, oral radiology, 117*(3), 343-352.

Downes, D. J., Chonofsky, M., Tan, K., Pfannenstiel, B. T., Reck-Peterson, S. L., & Todd, R. B. (2014). Characterization of the mutagenic spectrum of 4-nitroquinoline 1-oxide (4-NQO) in Aspergillus nidulans by whole genome sequencing. *G3: Genes, Genomes, Genetics, 4*(12), 2483-2492.

Du, S. S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R. R., & Singh, A. (2018). How many samples are needed to estimate a convolutional neural network? *Advances in Neural Information Processing Systems, 31*.

EI-Naggar, A. K. (2017). *WHO classification of head and neck tumours*: International Agency.

El Hallani, S., Poh, C., Macaulay, C., Follen, M., Guillaud, M., & Lane, P. (2013). Ex vivo confocal imaging with contrast agents for the detection of oral potentially malignant lesions. *Oral oncology, 49*(6), 582-590.

Farah, C. S., McIntosh, L., Georgiou, A., & McCullough, M. J. (2012). Efficacy of tissue autofluorescence imaging (VELScope) in the visualization of oral mucosal lesions. *Head Neck, 34*(6), 856-862.

Farahati, B., Stachs, O., Prall, F., Stave, J., Guthoff, R., Pau, H. W., & Just, T. (2010). Rigid confocal endoscopy for in vivo imaging of experimental oral squamous intra-epithelial lesions. *Journal of oral pathology & medicine, 39*(4), 318-327.

Feher, B., Tussie, C., & Giannobile, W. V. (2024). Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches. *Frontiers in Artificial Intelligence, 7*, 1427517.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., . . . Bray, F. (2020). Global cancer observatory: cancer today. *Lyon: International agency for research on cancer, 20182020*.

Fitzpatrick, S. G., Hirsch, S. A., & Gordon, S. C. (2014). The malignant transformation of oral lichen planus and oral lichenoid lesions: a systematic review. *The Journal of the American Dental Association, 145*(1), 45-56.

Fleischmann, T., Jirkof, P., Henke, J., Arras, M., & Cesarovic, N. (2016). Injection anaesthesia with fentanyl–midazolam–medetomidine in adult female mice: importance of antagonization and perioperative care. *Laboratory animals, 50*(4), 264-274.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics, 36*(4), 193-202.

Gadiwan, M., Madhushankari, G., Mandana, D., Praveen, S., Selvamani, M., & Pradeep, D. (2014). Nuclear features in different grades of epithelial dysplasia in leukoplakia: A computer assisted microscopic study. *Journal of Oral Maxillofacial Pathology, 18*(2), 194-200.

Gamarra, M., Zurek, E., San-Juan, H., Eng, S., & Norte, U. (2017). A study of image analysis algorithms for segmentation, feature extraction and classification of cells. *Journal of Information Systems Engineering Management, 2*(4), 20.

Gerondakis, S., Grumont, R., Gugasyan, R., Wong, L., Isomura, I., Ho, W., & Banerjee, A. (2006). Unravelling the complexities of the NF-κB signalling pathway using mouse knockout and transgenic models. *Oncogene, 25*(51), 6781-6799.

Goldsborough, T., Philps, B., O'Callaghan, A., Inglis, F., Leplat, L., Filby, A., . . . Bankhead, P. (2024). InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation. *arXiv, 15954*.

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health, 2*(1), e0000082.

Gonzalez-Moles, M., Scully, C., & Gil-Montoya, J. (2008). Oral lichen planus: controversies surrounding malignant transformation. *Oral Diseases, 14*(3), 229-243.

Goodrich, G. L., Bennett, R. R., De L'aune, W. R., Lauer, H., & Mowinski, L. (1979). Kurzweil reading machine: A partial evaluation of its optical character recognition error rate. *Journal of Visual Impairment Blindness, 73*(10), 389-399.

Grafton-Clarke, C., Chen, K. W., & Wilcock, J. (2019). Diagnosis and referral delays in primary care for oral squamous cell cancer: a systematic review. *British Journal of General Practice, 69*(679), e112-e126. doi:10.3399/bjgp18x700205

Güneri, P., & Epstein, J. B. (2014). Late stage diagnosis of oral cancer: Components and possible solutions. *Oral Oncology, 50*(12), 1131-1136. doi:https://doi.org/10.1016/j.oraloncology.2014.09.005

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine, 40*(2), 44-58.

Haxel, B. R., Goetz, M., Kiesslich, R., & Gosepath, J. (2010). Confocal endomicroscopy: a novel application for imaging of oral and oropharyngeal mucosa in human. *European archives of oto-rhino-laryngology, 267*(3), 443-448.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems their applications, 13*(4), 18-28.
Hellebust, A., Rosbach, K., Wu, J. K., Nguyen, J., Gillenwater, A. M., Vigneswaran, N., & Richards-Kortum, R. R. (2013). Vital-dye-enhanced multimodal imaging of neoplastic progression in a mouse model of oral carcinogenesis. *Journal of Biomedical Optics, 18*(12), 126017.
Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.
Holloway, L., Anees, A., Al Moiuee, D., Uddin, A., Haider, A., Gorse, D., . . . Bharathy, G. (2024). People RDC national pathfinder project: exploring federated learning tools, opportunities and resource requirements.
Holmstrup, P., Vedtofte, P., Reibel, J., & Stoltze, K. (2006). Long-term treatment outcome of oral premalignant lesions. *Oral Oncology, 42*(5), 461-474.
Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2020). *Explainable AI methods-a brief overview.* Paper presented at the International workshop on extending explainable AI beyond deep models and classifiers.
Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*: John Wiley & Sons.
Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., . . . Puliafito, C. A. (1991). Optical coherence tomography. *Science, 254*(5035), 1178-1181.
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*.
Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning? *arXiv, 1608.08614*.
Huisman, A., Ploeger, L. S., Dullens, H. F., Poulin, N., Grizzle, W. E., & van Diest, P. J. (2005). Development of 3D chromatin texture analysis using confocal laser scanning microscopy. *Analytical Cellular Pathology, 27*(5, 6), 335-345.
Humaira, H., & Rasyidah, R. (2020). *Determining the appropiate cluster number using elbow method for k-means algorithm.* Paper presented at the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA).
Hurlstone, D., Baraza, W., Brown, S., Thomson, M., Tiffin, N., & Cross, S. (2008). In vivo real-time confocal laser scanning endomicroscopic colonoscopy for the detection and characterization of colorectal neoplasia. *Journal of British Surgery, 95*(5), 636-645.
IARC. (2023). Oral cancer prevention. *IARC Handb Cancer Prev., 19*, 1-358.
Idrees, M., Halimi, R., Gadiraju, S., Frydrych, A. M., & Kujan, O. (2024). Clinical competency of dental health professionals and students in diagnosing oral mucosal lesions. *Oral Diseases, 30*(5), 3108-3116.
Iftikhar, S., & Davy, S. (2024). *Reducing Carbon Footprint in AI: A Framework for Sustainable Training of Large Language Models.* Paper presented at the Proceedings of the Future Technologies Conference.
Jabbour, J. M., Saldua, M. A., Bixler, J. N., & Maitland, K. C. (2012). Confocal endomicroscopy: instrumentation and medical applications. *Annals of biomedical engineering, 40*, 378-397.

Jabbour, J. M., Saldua, M. A., Bixler, J. N., & Maitland, K. C. (2012). Confocal Endomicroscopy: Instrumentation and Medical Applications. *Annals of Biomedical Engineering, 40*(2), 378-397. doi:10.1007/s10439-011-0426-y

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Unsupervised learning. In *An introduction to statistical learning: with applications in Python* (pp. 503-556): Springer.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior therapy, 51*(5), 675-687.

Johnson, O. V., Alyasiri, O. M., Akhtom, D. a., & Johnson, O. E. (2023). Image Analysis through the lens of ChatGPT-4. *Journal of Applied Artificial Intelligence, 4*(2), 31-46.

Jubair, F., Al-karadsheh, O., Malamos, D., Al Mahdi, S., Saad, Y., & Hassona, Y. (2022). A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Diseases, 28*(4), 1123-1130.

Kanojia, D., & Vaidya, M. M. (2006). 4-nitroquinoline-1-oxide induced experimental oral carcinogenesis. *Oral Oncology, 42*(7), 655-667.

Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. J. M. i. a. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis, 65*, 101759.

Keinänen, A., Uittamo, J., & Snäll, J. (2024). Do we recognize oral cancer? Primary professional delay in diagnosis of oral squamous cell carcinoma. *Clinical Oral Investigations, 28*(2), 131.

Kim, P. (2017). Matlab deep learning. *With machine learning, neural networks artificial intelligence, 130*(21), 151.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*: Springer.

Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., . . . Cuocolo, R. (2025). Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic interventional radiology, 31*(2), 75.

Koike, R., Uchiyama, T., Arimoto-Kobayashi, S., Okamoto, K., & Negishi, T. (2018). Increase of somatic cell mutations in oxidative damage-sensitive drosophila. *Genes Environment, 40*(1), 1-8.

Köntgen, F., Grumont, R. J., Strasser, A., Metcalf, D., Li, R., Tarlinton, D., & Gerondakis, S. (1995). Mice lacking the c-rel proto-oncogene exhibit defects in lymphocyte proliferation, humoral immunity, and interleukin-2 expression. *Genes development, 9*(16), 1965-1977.

Kraft, M., Glanz, H., von Gerlach, S., Wisweh, H., Lubatschowski, H., & Arens, C. (2008). Clinical value of optical coherence tomography in laryngology. *Head Neck: Journal for the Sciences Specialties of the Head*

30(12), 1628-1635.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*.

Kujan, O., Khattab, A., Oliver, R. J., Roberts, S. A., Thakker, N., & Sloan, P. (2007). Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation. *Oral Oncology, 43*(3), 224-231.

Kujan, O., Oliver, R. J., Khattab, A., Roberts, S. A., Thakker, N., & Sloan, P. (2006). Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. *Oral oncology, 42*(10), 987-993.

Kulwa, F., Li, C., Zhao, X., Cai, B., Xu, N., Qi, S., . . . Teng, Y. (2019). A state-of-the-art survey for microorganism image segmentation methods and future potential. *IEEE Access, 7*, 100243-100269.

Laitinen, A., & Sahlgren, O. (2021). AI systems and respect for human autonomy. *Frontiers in artificial intelligence, 4*, 151.

LeCun, Y., Bengio, Y., & Hinton, G. (2015a). Deep learning. *nature, 521*(7553), 436-444.

LeCun, Y., Bengio, Y., & Hinton, G. (2015b). Deep learning. *Nature methods, 521*(7553), 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Li, L., Fan, Y., Tse, M., & Lin, K.-Y. (2020). A review of applications in federated learning. *Computers Industrial Engineering, 149*, 106854.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks learning systems*.

Linxweiler, M., Al Kadah, B., Bozzato, A., Bozzato, V., Hasenfus, A., Kim, Y.-J., . . . Charalampaki, P. (2016a). Noninvasive histological imaging of head and neck squamous cell carcinomas using confocal laser endomicroscopy. *European archives of oto-rhino-laryngology, 273*(12), 4473-4483.

Linxweiler, M., Al Kadah, B., Bozzato, A., Bozzato, V., Hasenfus, A., Kim, Y.-J., . . . Charalampaki, P. (2016b). Noninvasive histological imaging of head and neck squamous cell carcinomas using confocal laser endomicroscopy. *European Archives of Oto-Rhino Laryngology, 273*(12), 4473-4483.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., . . . Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis, 42*, 60-88.

Liu, F., & Zhang, L. (2019). *View confusion feature learning for person re-identification.* Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision.

Lodi, G., Scully, C., Carrozzo, M., Griffiths, M., Sugerman, P. B., & Thongprasom, K. (2005). Current controversies in oral lichen planus: report of an international consensus meeting. Part 1. Viral infections and etiopathogenesis. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, Endodontology, 100*(1), 40-51.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., . . . Yang, H. (2024). Deepseek-vl: towards real-world vision-language understanding. *arXiv, 05525*.

Lucchese, A., Gentile, E., Romano, A., Maio, C., Laino, L., & Serpico, R. (2016a). The potential role of in vivo reflectance confocal microscopy for evaluating oral cavity lesions: a systematic review. *Journal of oral pathology & medicine, 45*(10), 723-729. doi:10.1111/jop.12454

Lucchese, A., Gentile, E., Romano, A., Maio, C., Laino, L., & Serpico, R. (2016b). The potential role of in vivo reflectance confocal microscopy for evaluating oral cavity lesions: a systematic review. *Journal of Oral Pathology Medicine, 45*(10), 723-729.

Luo, J., Young, C., Zhou, H., & Wang, X. (2018). Mouse models for studying oral cancer: impact in the era of cancer immunotherapy. *Journal of Dental Research, 97*(6), 683-690.
Lupu, M., Caruntu, A., Boda, D., & Caruntu, C. (2020). In Vivo Reflectance Confocal Microscopy-Diagnostic Criteria for Actinic Cheilitis and Squamous Cell Carcinoma of the Lip. *Journal of Clinical Medicine, 9*(6), 1987.
Lupu, M., Caruntu, A., Caruntu, C., Boda, D., Moraru, L., Voiculescu, V., & Bastian, A. (2018). Non-invasive imaging of actinic cheilitis and squamous cell carcinoma of the lip. *Molecular and Clinical Oncology, 8*(5), 640-646.
Lyu, L., Li, Y., Nandakumar, K., Yu, J., & Ma, X. (2020). How to democratise and protect AI: Fair and differentially private decentralised deep learning. *IEEE Transactions on Dependable Secure Computing, 19*(2), 1003-1017.
Ma, J., Schneider, L., Lapuschkin, S., Achtibat, R., Duchrau, M., Krois, J., . . . Samek, W. J. J. o. D. R. (2022). Towards trustworthy AI in dentistry. *Discovery!, 101*(11), 1263-1268.
MacEachern, S. J., & Forkert, N. D. (2021). Machine learning for precision medicine. *Genome, 64*(4), 416-425.
Maher, N., Collgros, H., Uribe, P., Ch'Ng, S., Rajadhyaksha, M., & Guitera, P. (2016). In vivo confocal microscopy for the oral cavity: Current state of the field and future potential. *Oral Oncology, 54*, 28-35.
Maher, N. G., Collgros, H., Uribe, P., Ch'Ng, S., Rajadhyaksha, M., & Guitera, P. (2016). In vivo confocal microscopy for the oral cavity: Current state of the field and future potential. *Oral oncology, 54*, 28-35. doi:10.1016/j.oraloncology.2016.01.003
Maitland, K. C., Gillenwater, A. M., Williams, M. D., El-Naggar, A. K., Descour, M. R., & Richards-Kortum, R. R. (2008). In vivo imaging of oral neoplasia using a miniaturized fiber optic confocal reflectance microscope. *Oral oncology, 44*(11), 1059-1066.
Marinov, D., & Karapetyan, D. (2019). *Hyperparameter optimisation with early termination of poor performers.* Paper presented at the 2019 11th Computer Science and Electronic Engineering (CEEC).
Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*: MIT press.
McCullough, M., Prasad, G., & Farah, C. (2010a). Oral mucosal malignancy and potentially malignant lesions: an update on the epidemiology, risk factors, diagnosis and management. *Australian dental journal, 55*, 61-65.
McCullough, M., Prasad, G., & Farah, C. J. A. d. j. (2010b). Oral mucosal malignancy and potentially malignant lesions: an update on the epidemiology, risk factors, diagnosis and management. *55*, 61-65.
McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for high performance scientific computing, 14*(9), 1-9.
Mocan, M. C., & Irkec, M. (2007). Fluorescein enhanced confocal microscopy in vivo for the evaluation of corneal epithelium. *Clinical Experimental Ophthalmology, 35*(1), 38-43.
Moore, C., Mehta, V., Ma, X., Chaudhery, S., Shi, R., Moore-Medlin, T., . . . Nathan, C. A. O. (2016). Interobserver agreement of confocal laser endomicroscopy for detection of head and neck neoplasia. *The Laryngoscope, 126*(3), 632-637.

Nag, R., & Das, R. K. (2018). Analysis of images for detection of oral epithelial dysplasia: A review. *Oral oncology, 78*, 8-15.
Nanci, A. (2017). Ten Cate's oral histology 9th Edition. Ninth edit. In: Elsevier Inc.
Nathan, C.-A. O., Kaskas, N. M., Ma, X., Chaudhery, S., Lian, T., Moore-Medlin, T., . . . Mehta, V. (2014). Confocal laser endomicroscopy in the detection of head and neck precancerous lesions. *Otolaryngology–Head and Neck Surgery, 151*(1), 73-80.
NHLBI, N. (2021). Quality assessment tool for observational cohort and cross-sectional studies. Retrieved from https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
NHMRC, N. H. a. M. R. C. (2019). Information paper: The implementation of the 3Rs in Australia. *NHMRC*.
Ni, Y., Yap, T., Silke, N., Silke, J., McCullough, M., Celentano, A., & O'Reilly, L. A. (2021). Loss of NF-kB1 and c-Rel accelerates oral carcinogenesis in mice. *Oral Diseases, 27*(2), 168-172.
Nikolenko, S. I. (2021). *Synthetic data for deep learning* (Vol. 174): Springer.
Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology, 24*(12), 1565-1567.
Odell, E., Kujan, O., Warnakulasuriya, S., & Sloan, P. (2021a). Oral epithelial dysplasia: recognition, grading and clinical significance. *Oral Diseases, 27*(8), 1947-1976.
Odell, E., Kujan, O., Warnakulasuriya, S., & Sloan, P. J. O. d. (2021b). Oral epithelial dysplasia: recognition, grading and clinical significance. *27*(8), 1947-1976.
Oetter, N., Knipfer, C., Rohde, M., von Wilmowsky, C., Maier, A., Brunner, K., . . . Stelzle, F. (2016). Development and validation of a classification and scoring system for the diagnosis of oral squamous cell carcinomas through confocal laser endomicroscopy. *Journal of translational medicine, 14*(1), 159.
Pachitariu, M., & Stringer, C. (2022). Cellpose 2.0: how to train your own model. *Nature methods, 19*(12), 1634-1641.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, 105906.
Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors, 52*(3), 381-410.
Park, W., Schwendicke, F., Krois, J., Huh, J.-K., & Lee, J.-H. (2023). Identification of dental implant systems using a large-scale multicenter data set. *Journal of Dental Research, 102*(7), 727-733.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017a). Automatic differentiation in pytorch.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017b). *Automatic differentiation in pytorch.* Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011a). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research, 12*, 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011b). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*, 2825-2830.

Peng, H., Shen, L., Zhou, G., & Wang, Y. (2020). Reflectance confocal microscopy characteristics of oral lichen planus: An analysis of 47 cases in a Chinese cohort. *Experimental and Therapeutic Medicine, 20*(5), 9134. doi:http://dx.doi.org/10.3892/ETM.2020.9134

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., . . . Wurbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biol, 18*(7), e3000410. doi:10.1371/journal.pbio.3000410

Phillips, C. J. (2020). Precision medicine and its imprecise history. *Harvard Data Science Review, 2*(1), 10.1162.

Pilonis, N. D., Januszewicz, W., & di Pietro, M. (2022). Confocal laser endomicroscopy in gastro-intestinal endoscopy: technical aspects and clinical applications. *Translational Gastroenterology Hepatology, 7*.

Pindborg, J. J., Reichart, P., Smith, C., & Van der Waal, I. (2012). *Histological typing of cancer and precancer of the oral mucosa: In collaboration with LH Sobin and Pathologists in 9 Countries*: Springer Science & Business Media.

Piorecka, K., Kurjata, J., & Stanczyk, W. A. (2022). Acriflavine, an acridine derivative for biomedical application: current state of the art. *Journal of Medicinal Chemistry, 65*(17), 11415-11432.

Pogorzelski, B., Hanenkamp, U., Goetz, M., Kiesslich, R., & Gosepath, J. (2012a). Systematic intraoperative application of confocal endomicroscopy for early detection and resection of squamous cell carcinoma of the head and neck: a preliminary report. *Archives of otolaryngology–head & neck surgery, 138*(4), 404-411.

Pogorzelski, B., Hanenkamp, U., Goetz, M., Kiesslich, R., & Gosepath, J. (2012b). Systematic intraoperative application of confocal endomicroscopy for early detection and resection of squamous cell carcinoma of the head and neck: a preliminary report. *Archives of otolaryngology–head neck surgery, 138*(4), 404-411.

Prieto, S. P., Powless, A. J., Boice, J. W., Sharma, S. G., & Muldoon, T. J. (2015). Proflavine hemisulfate as a fluorescent contrast agent for point-of-care cytology. *PloS one, 10*(5), e0125598.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining knowledge discovery, 9*(3), e1301.

Program, N. C. I.-S. E. R. (2022). Cancer stat facts: oral cavity and pharynx cancer. Retrieved from https://seercancergov/statfacts/html/oralcavhtml

Ragazzi, M., Piana, S., Longo, C., Castagnetti, F., Foroni, M., Ferrari, G., . . . Pellacani, G. (2014). Fluorescence confocal microscopy for pathologists. *Modern Pathology, 27*(3), 460-471.

Ramani, R. S., Tan, I., Bussau, L., Angel, C. M., McCullough, M., & Yap, T. (2022). Confocal microscopy in oral cancer and oral potentially malignant disorders: A systematic review. *Oral Diseases*.

Ramani, R. S., Tan, I., Bussau, L., Angel, C. M., McCullough, M., & Yap, T. (2023). Confocal microscopy in oral cancer and oral potentially malignant disorders: A systematic review. *Oral Diseases, 29*(8), 3003-3015.

Rangrez, M., Bussau, L., Ifrit, K., & Delaney, P. (2021). Fluorescence In Vivo Endomicroscopy: High-Resolution, 3-Dimensional Confocal Laser Endomicroscopy (Part 1). *Microscopy Today, 29*(2), 32-37.

Rangrez, M., Bussau, L., Ifrit, K., Preul, M. C., & Delaney, P. (2021). Fluorescence In vivo endomicroscopy part 2: applications of high-resolution, 3-dimensional confocal laser endomicroscopy. *Microscopy Today, 29*(3), 14-26.

Reibel, J., Gale, N., Hille, J., Hunt, J., Lingen, M., Muller, S., . . . Willams, M. (2017a). Oral potentially malignant disorders and oral epithelial dysplasia. *WHO classification of head neck tumours, 9*, 112.

Reibel, J., Gale, N., Hille, J., Hunt, J., Lingen, M., Muller, S., . . . Willams, M. (2017b). Oral potentially malignant disorders and oral epithelial dysplasia. In *WHO classification of head neck tumours* (4th ed., Vol. 9, pp. 112). Lyon: International Agency for Research on Cancer (IARC).

Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., . . . Antonelli, M. (2024). Understanding metric-related pitfalls in image analysis validation. *Nature methods, 21*(2), 182-194.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental science technology, 57*(9), 3464-3466.

Rischke, R., Schneider, L., Müller, K., Samek, W., Schwendicke, F., & Krois, J. (2022). Federated learning in dentistry: chances and challenges. *Journal of Dental Research, 101*(11), 1269-1273.

Robertson, T. A., Bunel, F., & Roberts, M. S. (2013). Fluorescein derivatives in intravital fluorescence imaging. *Cells, 2*(3), 591-606.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis machine intelligence, 32*(3), 569-575.

Roels, J., Aelterman, J., De Vylder, J., Lippens, S., Luong, H. Q., Guérin, C. J., & Philips, W. (2016). Image degradation in microscopic images: Avoidance, artifacts, and solutions. *Focus on bio-image informatics*, 41-67.

Rokhshad, R., Ducret, M., Chaurasia, A., Karteva, T., Radenkovic, M., Roganovic, J., . . . Lahoud, P. (2023). Ethical Considerations on Artificial Intelligence in Dentistry: A Framework and Checklist. *Journal of dentistry*, 104593.

Rokhshad, R., Mohammad-Rahimi, H., Price, J. B., Shoorgashti, R., Abbasiparashkouh, Z., Esmaeili, M., . . . Soltani, P. (2024). Artificial intelligence for classification and detection of oral mucosa lesions on photographs: a systematic review and meta-analysis. *Clinical Oral Investigations, 28*(1), 88.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.* Paper presented at the Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18.

Rutkowska, M., Hnitecka, S., Nahajowski, M., Dominiak, M., & Gerber, H. (2020). Oral cancer: The first symptoms and reasons for delaying correct diagnosis and appropriate treatment. *Advances in Clinical Experimental Medicine, 29*(6), 735-743.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., . . . Vedadi, E. (2024). Capabilities of gemini models in medicine. *arXiv, 18416*.

Sagheer, S. H., Whitaker-Menezes, D., Han, J. Y., Curry, J. M., Martinez-Outschoorn, U., & Philp, N. J. (2021). 4NQO induced carcinogenesis: A mouse model for oral squamous cell carcinoma. In *Methods in Cell Biology* (Vol. 163, pp. 93-111): Elsevier.

Saka-Herrán, C., Jané-Salas, E., Mari-Roig, A., Estrugo-Devesa, A., & López-López, J. (2021). Time-to-Treatment in Oral Cancer: Causes and Implications for Survival. *Cancers, 13*(6), 1321. doi:10.3390/cancers13061321

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Schmid, B. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods, 9*(7), 676-682.

Schindelin, J., Rueden, C. T., Hiner, M. C., & Eliceiri, K. W. (2015). The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular reproduction development, 82*(7-8), 518-529.

Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). *Cell detection with star-convex polygons.* Paper presented at the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11.

Schmied, C., Nelson, M. S., Avilov, S., Bakker, G.-J., Bertocchi, C., Bischof, J., . . . Chiritescu, C. J. N. M. (2024). Community-developed checklists for publishing images and image analyses. *21*(2), 170-181.

Schneider, L., Rischke, R., Krois, J., Krasowski, A., Büttner, M., Mohammad-Rahimi, H., . . . Uribe, S. E. (2023). Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs. *Journal of dentistry, 135*, 104556.

Schwendicke, F., Singh, T., Lee, J.-H., Gaudin, R., Chaurasia, A., Wiegand, T., . . . Krois, J. (2021). Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *Journal of Dentistry, 107*, 103610.

Seoane, J., Alvarez-Novoa, P., Gomez, I., Takkouche, B., Diz, P., Warnakulasiruya, S., . . . Varela-Centelles, P. (2016). Early oral cancer diagnosis: The Aarhus statement perspective. A systematic review and meta-analysis. *Head & Neck, 38*(S1), E2182-E2189. doi:10.1002/hed.24050

Shamim, M. Z. M., Syed, S., Shiblee, M., Usman, M., Ali, S. J., Hussein, H. S., & Farrag, M. (2022). Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. *The Computer Journal, 65*(1), 91-104.

Shavlokhova, V., Flechtenmacher, C., Sandhu, S., Pilz, M., Vollmer, M., Hoffmann, J., . . . Freudlsperger, C. (2020). Detection of oral squamous cell carcinoma with ex vivo fluorescence confocal microscopy: sensitivity and specificity compared to histopathology. *Journal of Biophotonics, 13*, e202000100.

Shavlokhova, V., Flechtenmacher, C., Sandhu, S., Vollmer, M., Hoffmann, J., Engel, M., . . . Freudlsperger, C. (2021). Features of oral squamous cell carcinoma in ex vivo fluorescence confocal microscopy. *International journal of dermatology, 60*(2), 236-240. doi:https://dx.doi.org/10.1111/ijd.15152

Shavlokhova, V., Flechtenmacher, C., Sandhu, S., Vollmer, M., Vollmer, A., Pilz, M., . . . Freudlsperger, C. (2021). Feasibility and Implementation of Ex Vivo Fluorescence Confocal Microscopy for Diagnosis of Oral Leukoplakia: Preliminary Study. *Diagnostics (Basel, Switzerland), 11*(6). doi:https://dx.doi.org/10.3390/diagnostics11060951

Shavlokhova, V., Sandhu, S., Flechtenmacher, C., Koveshazi, I., Neumeier, F., Padrón-Laso, V., . . . Vollmer, A. (2021). Deep learning on oral squamous cell carcinoma ex vivo fluorescent confocal microscopy data: a feasibility study. *Journal of Clinical Medicine, 10*(22), 5326.

Shinohara, S., Funabiki, K., Kikuchi, M., Takebayashi, S., Hamaguchi, K., Hara, S., . . . Mizoguchi, A. (2020). Real-time imaging of head and neck squamous cell carcinomas using confocal micro-endoscopy and applicable dye: A preliminary study. *Auris Nasus Larynx, 47*(4), 668-675.

Shneiderman, B. (2020). Design lessons from AI's two grand goals: human emulation and useful applications. *IEEE Transactions on Technology Society, 1*(2), 73-82.

Sieracki, M. E., Reichenbach, S. E., & Webb, K. L. (1989). Evaluation of automated threshold selection methods for accurately sizing microscopic fluorescent cells by image analysis. *Applied Environmental microbiology, 55*(11), 2762-2772.

Sievert, M., Stelzle, F., Aubreville, M., Mueller, S. K., Eckstein, M., Oetter, N., . . . Goncalves, M. (2021). Intraoperative free margins assessment of oropharyngeal squamous cell carcinoma with confocal laser endomicroscopy: a pilot study. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery, 278*(11), 4433-4439. doi:https://dx.doi.org/10.1007/s00405-021-06659-y

Speight, P. M., Khurram, S. A., & Kujan, O. (2018a). Oral potentially malignant disorders: risk of progression to malignancy. *Oral surgery, oral medicine, oral pathology oral radiology, 125*(6), 612-627.

Speight, P. M., Khurram, S. A., & Kujan, O. (2018b). Oral potentially malignant disorders: risk of progression to malignancy. *Oral surgery, oral medicine, oral pathology, oral radiology, 125*(6), 612-627.

Standring, S. (2020). *Gray's Anatomy, 42nd Edition*. Amsterdam: Elsevier.

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., & Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. *BMC bioinformatics, 22*, 1-11.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B, 36*(2), 111-133.

Stringer, C., & Pachitariu, M. (2025). Cellpose3: one-click image restoration for improved cellular segmentation. *Nature methods*, 1-8.

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature methods, 18*(1), 100-106.

Strubell, E., Ganesh, A., & McCallum, A. (2020). *Energy and policy considerations for modern deep learning research.* Paper presented at the Proceedings of the AAAI conference on artificial intelligence.

Sun, M.-L., Liu, Y., Liu, G., Cui, D., Heidari, A. A., Jia, W.-Y., . . . Luo, Y. (2020). Application of machine learning to stomatology: a comprehensive review. *IEEE Access, 8*, 184360-184374.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Tan, J., Quinn, M., Pyman, J., Delaney, P., & McLaren, W. (2009). Detection of cervical intraepithelial neoplasia in vivo using confocal endomicroscopy. *BJOG: An International Journal of Obstetrics Gynaecology, 116*(12), 1663-1670.

Tan, M., & Le, Q. (2019). *Efficientnet: Rethinking model scaling for convolutional neural networks.* Paper presented at the International conference on machine learning.

Tanriver, G., Soluk Tekkesin, M., & Ergen, O. (2021). Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers, 13*(11), 2766.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513*, 429-441.

Tsantoulis, P., Kastrinakis, N., Tourvas, A., Laskaris, G., & Gorgoulis, V. (2007). Advances in the biology of oral cancer. *Oral Oncology, 43*(6), 523-534.

Tubbs, R. K., Ditmars Jr, W. E., & Van Winkle, Q. (1964). Heterogeneity of the interaction of DNA with acriflavine. *Journal of Molecular Biology, 9*(2), 545-557.

Ulrich, M., González, S., Lange-Asschenfeldt, B., Roewert-Huber, J., Sterry, W., Stockfleth, E., & Astner, S. (2011a). Non-invasive diagnosis and monitoring of actinic cheilitis with reflectance confocal microscopy. *Journal of the European Academy of Dermatology Venereology, 25*(3), 276-284.

Ulrich, M., González, S., Lange-Asschenfeldt, B., Roewert-Huber, J., Sterry, W., Stockfleth, E., & Astner, S. (2011b). Non-invasive diagnosis and monitoring of actinic cheilitis with reflectance confocal microscopy. *Journal of the European Academy of Dermatology and Venereology, 25*(3), 276-284.

Uttam, S., Pham, H. V., LaFace, J., Leibowitz, B., Yu, J., Brand, R. E., . . . Liu, Y. (2015). Early prediction of cancer progression by depth-resolved nanoscale mapping of nuclear architecture from unstained tissue specimens. *Cancer research, 75*(22), 4718-4727.

Van der Meij, E., Mast, H., & van der Waal, I. (2007). The possible premalignant character of oral lichen planus and oral lichenoid lesions: a prospective five-year follow-up study of 192 patients. *Oral Oncology, 43*(8), 742-748.

Van der Meij, E., & Van der Waal, I. (2003). Lack of clinicopathologic correlation in the diagnosis of oral lichen planus based on the presently available diagnostic criteria and suggestions for modifications. *Journal of oral pathology medicine, 32*(9), 507-512.

Van der Waal, I. (2009). Oral lichen planus and oral lichenoid lesions; a critical appraisal with emphasis on the diagnostic aspects. *Medicina oral, patologia oral y cirugia bucal, 14*(7), E310-E314.

Van der Waal, I. (2013). Are we able to reduce the mortality and morbidity of oral cancer; some considerations. *Medicina oral, patologia oral y cirugia bucal, 18*(1), e33.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in science engineering, 13*(2), 22-30.

Van Rossum, G., & Drake, F. (2021). Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009. Retrieved from https://www.python.org
van Rossum, G., & Drake, F. (2021). Python Language Reference (Release 3.7. 10). In: Python Software Foundation.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods, 17*(3), 261-272.
Volgger, V., Conderman, C., & Betz, C. S. (2013a). Confocal laser endomicroscopy in head and neck cancer. *Current Opinion in Otolaryngology & Head and Neck Surgery, 21*(2), 164-170. doi:10.1097/moo.0b013e32835df135
Volgger, V., Conderman, C., & Betz, C. S. (2013b). Confocal laser endomicroscopy in head and neck cancer: steps forward? *Current opinion in otolaryngology head neck surgery, 21*(2), 164-170.
von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., . . . Karinou, E. (2021). Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nature communications, 12*(1), 2276.
Wang, T. D., Mandella, M. J., Contag, C. H., & Kino, G. S. (2003). Dual-axis confocal microscope for high-resolution in vivo imaging. *Optics letters, 28*(6), 414-416.
Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., . . . Chau, D. H. P. (2020). CNN explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization Computer Graphics, 27*(2), 1396-1406.
Warnakulasuriya, S. (2020a). Oral potentially malignant disorders: A comprehensive review on clinical aspects and management. *Oral Oncology, 102*, 104550.
Warnakulasuriya, S. (2020b). *Textbook of Oral Cancer: Prevention, Diagnosis Management*.
Warnakulasuriya, S., Kujan, O., Aguirre-Urizar, J. M., Bagan, J. V., González-Moles, M. Á., Kerr, A. R., . . . Ogden, G. R. (2021). Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer. *Oral Diseases, 27*(8), 1862-1880.
Warnakulasuriya, S., Kujan, O., Aguirre-Urizar, J. M., Bagan, J. V., González-Moles, M. Á., Kerr, A. R., . . . Johnson, N. W. (2021). Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer. *Oral Diseases, 27*(8), 1862-1880. doi:10.1111/odi.13704
Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020). *Star-convex polyhedra for 3D object detection and segmentation in microscopy.* Paper presented at the Proceedings of the IEEE/CVF winter conference on applications of computer vision.
Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data, 3*(1), 1-40.
Welikala, R. A., Remagnino, P., Lim, J. H., Chan, C. S., Rajendran, S., Kallarakkal, T. G., . . . Kerr, A. R. (2020). Automated detection and classification of oral lesions

using deep learning for early detection of oral cancer. *IEEE Access, 8*, 132677-132693.

WHO. (2013). *International Classification of Diseases for Oncology* (3rd ed.): World Health Organization.

Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., . . . Antani, S. K. (2023). Assessing inter-annotator agreement for medical image segmentation. *IEEE Access, 11*, 21300-21312.

Yap, T., Tan, I., Ramani, R. S., Bhatia, N., Demetrio de Souza Franca, P., Angel, C., . . . McCullough, M. J. (2023). Acquisition and annotation in high resolution in vivo digital biopsy by confocal microscopy for diagnosis in oral precancer and cancer. *Frontiers in Oncology, 13*, 1209261.

Yeladandi, M., Sundaram, U. T. N., Muthukumaran, D., Umamaheswari, T., & Dhanya, M. (2024). A Cross-Sectional Study on Oral Potentially Malignant Disorders: Diagnostic Challenges in Early Detection of Dysplasia and the Role of Velscope. *Cureus, 16*(9).

Zhang, H., & Qie, Y. (2023). Applying deep learning to medical imaging: a review. *Applied Sciences, 13*(18), 10521.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., . . . He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43-76.

11. APPENDICES

11.1. Appendix 1 – Search terms for systematic review in Chapter 2

Article search terms for this review:

"Microscopy, Confocal" or "confocal" or "Confocal microscop*" or "reflectance confocal microscop*" or "in vivo microscop*" or "endomicroscop*") AND ("oral" or "mouth" or "orophar*" or "oropharynx*" or "throat*" or "tongue" or "floor of mouth" or "palate" or "lingual" or "buccal" or "lip" or "labial" or "tonsil*" or "mucosa*" or "retromolar" or "cheek" or "gingiva" or "intra-oral" or "vermillion border") AND ("pre-cancer*" or "cancer*" or "neoplasm*" or "malignan*" or "tumour" or "tumor" or "neoplasia" or "carcinoma" or "Leukoplakia" or "white patches" or "erythroplakia" or "red patches" or "erythroleukoplakia" or "precancer*" or "oral potentially malignant disorder" or "oral potentially malignant lesion" or "proliferative verrucous leukoplakia" or "dysplasia" or "premalignant" or "pre-malignant" or "lichen planus" or "oral submucous fibrosis" or "oral lupus erythematosus" or "actinic cheilitis" or "lichenoid" or "hyperplas*" or "ulcer" or "nodul*" or "metasta*" or "pemphigus vulgaris" OR "pemphigoid" OR "autoimmune bullous disease*" OR "autoimmune bullous disorder*" OR "immunobullous disease*" or "melanoma" or "macule" or "naev*"

11.2. Appendix 2 – All programming code developed

The complete code, data processing scripts, and experiment configurations used in this thesis are available in the following GitHub repository:

Web address link: **https://github.com/Rira-zen/Machine_learning_digital_biopsies_oral_cancer/tree/main**

The following files can be found in the repository called '**Machine_learning_digital_biopsies_oral_cancer**':

1. **cnn_train_eval.py** = Python script for training CNN models using hyperparameter optimisation and cross validation
2. **chi_squared_ML_train_eval.py** = Python script for chi-squared statistical tests and development of the logistic regression, SVM, random forest, and XGBoost models using hyperparameter optimisation and cross validation on the data from excel worksheets
3. **stardist_ZeroCostDL4Mic.ipynb** = Python Jupyter notebook script for StarDist 2D segmentation model training and evaluation developed by (von Chamier et al., 2021) based on the work of (Schmidt et al., 2018).
4. **nuclei_clustering_measurements.py** = Clustering all nucleus measurements for different values of k, z-score standardisation, and image feature vector construction to prepare the data for ML analysis
5. **ANOVA_Tukey_stats.py** = Summary statistics of mean, median, and standard deviation along with analysis of variance and tukey's pairwise comparison post hoc test for nuclei measurements

To reproduce the experiments, clone the repository, install the dependencies listed in 'readme.txt' and run the scripts from the repository after modified them based on the use case.