University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Doig, KD;Perera, R;Kankanige, Y;Fellowes, A;Li, J;Lupat, R;Thompson, ER;Blombery, P;Fox, SB

Title:
Using artificial intelligence (AI) to model clinical variant reporting for next generation sequencing (NGS) oncology assays

Date:
2025

Citation:
Doig, K. D., Perera, R., Kankanige, Y., Fellowes, A., Li, J., Lupat, R., Thompson, E. R., Blombery, P. & Fox, S. B. (2025). Using artificial intelligence (AI) to model clinical variant reporting for next generation sequencing (NGS) oncology assays. BioData Mining, 18 (1), https://doi.org/10.1186/s13040-025-00489-y.

Persistent Link:
https://hdl.handle.net/11343/367778

License:
CC BY-NC-ND

Doig *et al. BioData Mining* (2025) 18:74
https://doi.org/10.1186/s13040-025-00489-y

BioData Mining

RESEARCH Open Access

Using artificial intelligence (AI) to model clinical variant reporting for next generation sequencing (NGS) oncology assays

Kenneth D. Doig[1,2,3*], Rashindrie Perera[1], Yamuna Kankanige[2,3], Andrew Fellowes[2], Jason Li[1], Richard Lupat[1], Ella R. Thompson[2,3], Piers Blombery[2] and Stephen B. Fox[2,3,4]

*Correspondence:
Kenneth D. Doig
ken.doig@petermac.org
[1]Research Division, Peter MacCallum Cancer Centre, Parkville, VIC, Australia
[2]Department of Pathology, Peter MacCallum Cancer Centre, Parkville, VIC, Australia
[3]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[4]Department of Pathology, University of Melbourne, Melbourne , Australia

Abstract

Background Targeted next generation sequencing (NGS) of somatic DNA is now routinely used for diagnostic and predictive reporting in the oncology clinic. The expert genomic analysis required for NGS assays remains a bottleneck to scaling the volume of patients being assessed. This study harnesses data from targeted clinical sequencing to build machine learning models that predict whether patient variants should be reported.

Methods Three somatic assays were used to build machine learning prediction models using the estimators Logistic Regression, Random Forest, XGBoost and Neural Networks. Using manual expert curation to select reportable variants as ground truth, we built models to classify clinically reportable variants. Assays were performed between 2020 and 2023 yielding 1,350,018 variants and used to report on 10,116 patients. All variants, together with 211 annotations and sequencing features, were used by the models to predict the likelihood of variants being reported.

Results The tree-based ensemble models performed consistently well achieving between 0.904 and 0.996 on the precision recall/area under the curve (PRC AUC) metric when predicting whether a variant should be reported. To assist model explainability, individual model predictions were presented to users within a tertiary analysis platform as a waterfall plot showing individual feature contributions and their values for the variant. Over 30% of the model performance was due to features sourced from statistics derived in-house from the sequencing assay precluding easy generalization of the models to other assays or other laboratories.

Conclusions Longitudinally acquired NGS assay data provide a strong basis for machine learning models for decision support to select variants for clinical oncology reports. The models provide a framework for consistent reporting practices and reducing inter-reviewer variability. To improve model transparency, individual variant predictions are able to be presented as part of reviewer workflows.

Keywords Targeted sequencing, Clinical diagnostics, Variant calling, Somatic mutations, Cancer genomics, Precision oncology, Machine learning, AI prediction algorithms, CDSS, Clinical decision support systems

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Clinical diagnostics in oncology has harnessed Next Generation Sequencing (NGS) to analyse patient DNA at the nucleotide level. This has allowed identification of diagnostic, prognostic and therapeutic information critical to a patient's management. In earlier work [1–3], we have highlighted limitations in scaling clinical sequencing, in particular, the effort required for the expert curation of the large number of somatic variants observed in a cancer patient's DNA. Within a clinical laboratory, despite improvements in tertiary analysis software and databases, qualitative decisions about variants are frequently required by standard operating procedures (SOPs) and many subjective choices must be made while compiling a routine NGS clinical report. This often leads to inter-reviewer variability or the risk of clinically significant variants not being reported [4–7]. This study harnesses a corpus of over one million variants sequenced from 10,116 patients to build a number of AI models to support genomic scientists in identifying the variants that warrant curation. By using the expert curation and clinical reporting of 18,841 variants, we have built machine learning models that provide decision support tools for expert curators and provide objective, quantitative metrics to assess patient variants.

Methods

NGS reporting workflow

This study used clinical patient data collected to provide diagnostic oncology reports. Samples for each assay were received and processed according to standard laboratory procedures and then sequenced. An automated bioinformatics pipeline was triggered at the completion of sequencing, which processed and annotated the sequencing data before loading the results into a tertiary analysis platform. Trained genomic scientists then analyse the variants and, based on quality control (QC) metrics and a large number of annotations, selected appropriate variants for reporting. Detailed descriptions of laboratory processes have been described previously [3] and reviewed more recently [2].

Sequenced variants were uploaded to either an in-house tertiary analysis decision support software system called PathOS [3] ('Haem' and 'Myeloid' assays) or a commercial analysis platform, PierianDX CGW (Solid Tumour), for filtering, analysis and reporting. Reported variants were manually curated to establish variant action within a patient's clinical context using either ACMG or AMP guidelines or other custom classifications [8, 9]. Curated variants with enriched expert annotations were deposited within a common database enabling subsequent patients presenting with the same variants to be matched to the existing variant annotations so that prospectively; only novel variants need be curated. The patient's clinical context was also stored with curated variants to inform decisions on whether the same variant appearing in a different clinical context warrant using the same stored curation or whether a new distinct, and perhaps adapted, curation of the variant and context would be required. For details of the pipelines and curation workflows please refer to previously documented processes [1–3] and the Supplementary Methods section.

Data extraction

This study used clinical patient data collected over a period of four years within the Pathology Department at the Peter MacCallum Cancer Centre (see Table 1).

Table 1 NGS assays characteristics used in study

Abbreviation	Assay	Runs	Patients	Total Variants	Unique Variants	Reported Variants	Unique Reportable	Reportable Genes
TSO500	Illumina TruSight Oncology 500	195	1,772	804,074	33,809	5,112	2,514	249
Haem	QIAGEN QIAseq custom panel	157	5,815	444,073	20,126	11,816	4,554	57
Myeloid	QIAGEN QIAseq custom panel (filtered gene list)	157	2,529	101,871	4,079	1,913	774	8 (Oct21-Jun22) or 22 (Jul22-Dec23)
			10,116	1,350,018		18,841		

Variant data was extracted from tertiary analysis systems (PathOS or PierianDX CGW), and a number of other annotation sources were used to enrich the data using data source APIs (MyVariant, Genome Nexus) or downloaded datasets (Cosmic Cancer Mutation Census [10], DeepMind AlphaMissense [11], MSK Cancer Hotspots [12, 13]). The annotation sources were matched to the tertiary analysis data using the GRCh37 genomic representation of the variant (chromosome, position, reference bases, alternate bases). Both the MyVariant and Genome Nexus sources aggregate multiple third-party data sources providing annotation efficiencies. MyVariant provided access to the sources dbNSFP [14], CADD [15], CIViC [16], ClinVar [17] and gnomAD [18] while Genome Nexus was used to access OncoKB [19], SignalDB [20] and dbSNP. To match downloaded Cosmic Mutation Census data, the sequenced dataset were queried by the HGNC gene and HGVS coding (HGVSc) format of the variant nomenclature.

The resulting annotated variant set used for the machine learning ingestion process comprised 1,350,018 rows (variants) and 211 columns (features). The split between the three project assays is shown in Table 1. The pruned subset of annotations (features) used for the models is described in Table 2.

To emulate the steps used by many laboratories to enrich variants for relevant somatic mutations, the following filters were applied to the full dataset.

1. Removal of variants with predicted consequences (synonymous, intronic, 5' UTR) unless they also were predicted to be a splice variant (remove variants without protein coding consequences),
2. Removal of variants with a variant frequency less than 2% (remove variants occurring towards the assay limits of detection),

Table 2 Feature set used for final Models. List of all features used for final models with in-house features highlighted in orange. All features are numerical except for categorical features marked "(cat)". The features have been categorised into four groups; sequencing and assay specific features (Seq), curated knowledge bases, both public and in-house (KB), reference genome variant effect (G) and public in-silico predictors (IS)

Feature Name	Feature Group	Source	Description	Source Organisation	Research Only License	URL
var_freq	Seq	PathOS	Variant allele frequency.	Peter MacCallum Cancer Centre	No	https://www.petermac.org/
var_depth	Seq	PathOS	Variant allele read depth.	Peter MacCallum Cancer Centre	No	https://www.petermac.org/
curated_class (cat)	KB	PathOS,PierianDX	Previously curated variant significance (PathOS categories: Clinically Significant, Not Clinically Significant, Unclear Clinically Significance, Unclassified) (PierianDX categories: Tier I-IV)	Peter MacCallum Cancer Centre, PierianDX	No	https://www.petermac.org/ https://www.pieriandx.com/
var_panel_pct	Seq	PathOS	Variant frequency for all samples in assay.	Peter MacCallum Cancer Centre	No	https://www.petermac.org/
Hotspot(cat)	KB	Cancer Hotspots	Variant located at recurrent cancer hotspot (0,1)	Memorial Sloan Kettering	No	www.cancerhotspots.org
okb_oncogenic (cat)	KB	Genome Nexus	OncoKB Oncogenic class (Oncogenic, Likely Oncogenic, Likely Neutral, Resistance, Inconclusive)	Memorial Sloan Kettering	Yes	www.oncokb.org
alpha_predict	IS	AlphaMissense	Missense variant pathogenicity prediction from protein structure and evolutionary conservation.	Google DeepMind Technologies	Yes	console.cloud.google.com/storage/browser/dm_alphamissense
okb_fda_level (cat)	KB	Genome Nexus	OncoKB FDA Approval Level (LEVEL_FDA2, LEVEL_FDA3)	Memorial Sloan Kettering	Yes	www.oncokb.org
cadd_rawscore	IS	MyVariant	Combined Annotation Dependent Depletion (CADD) raw scores	University of Washington	Yes	cadd.gs.washington.edu
cmc_mut_pct	KB	Cosmic Cancer Mutation Census	Cosmic Cancer Mutation Census (CMC) Percent of samples containing mutation.	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
dbnsfp_bayesdel_add_af_rankscore	IS	MyVariant.dbnsfp	Deleteriousness predictor rank score	University of Utah	No	fenglab.chpc.utah.edu/BayesDel
pop_pct	KB	MyVariant	Gnomad genome or exome population allele frequency	Broad Institute	No	gnomad.broadinstitute.org
cmc_gerp	IS	Cosmic Cancer Mutation Census	CMC Genomic Evolutionary Rate Profiling (GERP) score	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
signal_mutStatus (cat)	KB	Genome Nexus	Somatic status in Signal Database (germline, somatic)	Memorial Sloan Kettering	No	www.signaldb.org
cmc_mut_tier(cat)	KB	Cosmic Cancer Mutation Census	CMC Mutation Tier (0,1,2,3)	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
cmc_aa_type (cat)	G	Cosmic Cancer Mutation Census	CMC Amino Acid Mutation Type (Substitution - Nonsense/Missense/Coding silent, Insertion - Frameshift/In frame, Deletion – Frameshift/In frame, Complex – frameshift)	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
consequence (cat)	G	VEP	Variant Effect Predictor (VEP) calculated variant consequences	Ensembl	No	www.ensembl.org
con_ben	G	VEP	Summarisation of 'consequence' feature with single most deleterious consequence per variant.	Ensembl	No	www.ensembl.org
dbnsfp_primateai_rankscore	IS	MyVariant.dbnsfp	Missense variant pathogenicity prediction from protein structure and evolutionary conservation, rank score.	Illumina	No	doi.org/10.1038/s41588-018-0167-z
dbnsfp_sift4g_converted_rankscore	IS	MyVariant.dbnsfp	SIFT 4G missense pathogenicity prediction converted rank score	Agency for Science, Technology and Research	No	www.nature.com/nprot/journal/v11/n1/full/nprot.2015.123.html
dbnsfp_fathmm_mkl_coding_rankscore	IS	MyVariant.dbnsfp	Functional Analysis through Hidden Markov Models, Multiple Kernel Learning variant functional prediction coding rank score	University of Bristol	No	doi:10.1093/bioinformatics/btv009
var_sam_tot	Seq	PathOS	Count of variants in sequenced sample.	Peter MacCallum Cancer Centre	No	https://www.petermac.org/
exonpct	G	VEP	Exon position as a percentage in transcript	Ensembl	No	www.ensembl.org
vartype (cat)	G	Genome Nexus	dbSNP variant type (SNP,DEL,INS,DNP,TNP,ONP)	NCBI	No	www.ncbi.nlm.nih.gov/snp
dbnsfp_bstatistic_converted_rankscore	IS	MyVariant.dbnsfp	Primate variant selection predictor rank score	University of Washington	No	DOI: 10.1371/journal.pgen.1000471
ndbnsp	KB	VEP	Number of dbsnp entries for variant	Ensembl	No	www.ensembl.org
dbnsfp_genocanyon_rankscore	IS	MyVariant.dbnsfp	GenoCanyon variant functional predictor rank score	Yale School of Public Health	No	DOI: 10.1038/srep10576
dbnsfp_clinvar_sig (cat)	KB	MyVariant.dbnsfp	Clinvar clinical significance (Pathogenic,Likely_pathogenic,Benign, Likely_benign,Conflicting)	NCBI	Yes	www.ncbi.nlm.nih.gov/clinvar
tsg(cat)	KB	Cosmic Cancer Mutation Census	CMC Tumour Suppressor Gene (0,1)	Sanger Institute	Yes	cancer.sanger.ac.uk/census
cmc_onc(cat)	KB	Cosmic Cancer Mutation Census	CMC Oncogene (0,1)	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
npubmed	KB	VEP	Number of Pubmed articles for variant	Ensembl	No	www.ensembl.org
signal_pathogenic(cat)	KB	Genome Nexus	Marked pathogenic in Signal Database (0,1)	Memorial Sloan Kettering	No	www.signaldb.org
cmc_cgc_tier(cat)	KB	Cosmic Cancer Mutation Census	CMC Gene Tier (1,2)	Sanger Institute	Yes	cancer.sanger.ac.uk/cmc
single_umi	Seq	PathOS	Percentage of reads containing singleton UMI (Unique Molecular Identifier) – assay QC metric (not used in TSO500 assay).	Peter MacCallum Cancer Centre	No	https://www.petermac.org/
fr_balance	Seq	PathOS	Percentage balance between forward and reverse reads containing variant – assay QC metric (not used in TSO500 assay).	Peter MacCallum Cancer Centre	No	https://www.petermac.org/

3. Removal of variants occurring at greater that 1% in the population (frequently occurring variants are unlikely to be deleterious),
4. Removal of variants occurring in more than 35% of the samples within the assay (frequently occurring variants are likely to be assay artifacts or frequently occurring in the population).

Both the full set of variants and the filtered variant dataset were used to build and compare machine learning models.

To compare the relative importance of features by assay, the scikit-learn Random Forest feature importance property was used which provides an impurity-based score per feature summing to 1. For categorical features, individual category values within a feature were summed to give a single feature value.

Machine learning

A schematic of the basic ML workflow is shown in Figure S3 Workflow.

Data ingestion and pre-processing

Data processing was employed to ensure suitability for model fitting. The raw data was filtered to identify numerical and categorical features. Uninformative feature columns were removed from the data set, including HGVS formatted variants, patient identifiers, database accession numbers, reporting data entered after genomic analysis and free text descriptors. This step reduced the initial raw datasets from 211 to 157 features. Two separate preprocessor pipelines were then employed to handle the two different data types. For numerical data, missing values were imputed with a constant value of 0. Scikit-learn's RobustScaler was then used to normalise these features. RobustScaler removes the median and scales the data according to the quantile range, effectively reducing the influence of outliers on the model's learning process. Missing values in categorical features were imputed with a dummy value of 'Other' which did not occur within categorical features. It was not appropriate to drop rows with missing values as the annotation databases used to enrich the variant features often had limited coverage of the observed sequence variants. Instead, For the major annotation data sources (Cosmic CMC, MyVariant, Genome Nexus, HotSpot and AlphaMissense) additional binary features were added indicating missing data for the rows (incosmic, inmyvariant, ingenomenexus, inhotspot, inalphamissense). The addition of these 'data exists' binary features allow the models to differentiate between the absence of data and imputed data. The imputed categorical data was then converted into numerical representations using Scikit-learn's OneHotEncoder. This encoder replaces categorical features with a set of binary features, one for each distinct category value.

Model estimators

To explore a range of learning algorithms and identify the most effective models for predicting variant reporting, four diverse machine learning models were employed:

- Logistic Regression Classifier (LR): This model implements regularized logistic regression and can handle both dense and sparse input.
- Random Forest (RF): This tree-based ensemble method combines multiple decision trees, where each tree makes a prediction, and the final prediction is the majority vote from all trees. RF is known for its robustness to noise and ability to handle complex relationships within the data.
- Extreme Gradient Boosting (XGB): This powerful tree-based ensemble method leverages gradient boosting to iteratively learn from the training data, focusing on areas where previous models made errors. XGBoost is known for its accuracy and ability to handle complex features.
- Neural Network (NN): This flexible model learns complex, non-linear relationships between features and the target variable. A multi-layer perceptron architecture with backpropagation was used for training. NNs can be particularly effective in capturing intricate patterns within the data.

By utilizing this set of models, a broad range of learning styles were captured to identify the approach that generalises best to unseen data for accurate variant clinical reporting prediction.

Model estimator hyperparameters and architectures

The parameters and the architecture of the models were chosen empirically. The LR classifier employed default hyperparameters, while the Random Forest classifier was configured with 300 decision trees (n_estimators = 300) to create a robust ensemble model. The XGBoost regressor was configured with 1,000 estimators (n_estimators = 1000) and a maximum depth of 20 (max_depth = 20) to enhance its ability to capture complex relationships within the data. All three models utilised a constant random state to ensure consistent results across training runs.

A multi-layer perceptron architecture was implemented for the NN model. The architecture utilises a sequential stack of densely connected layers with ReLU (Rectified Linear Unit) activation functions for non-linearity. The first hidden layer has 50 neurons, followed by two additional hidden layers with 6 neurons each. Dropout layers with a rate of 0.4 are incorporated after the first two hidden layers to minimise overfitting. The final output layer has a single neuron with a sigmoid activation function, generating a probability value between 0 and 1 representing the predicted likelihood of a variant requiring clinical reporting. The Adam optimiser was used for efficient training, and the binary cross-entropy loss function was employed to measure prediction error during the training process. Additionally, early stopping with a patience of 1 was employed to stop training once the validation accuracy reaches the maximum to avoid overfitting on the training data.

Feature pruning

To identify the most informative features for variant clinical reporting prediction, recursive feature elimination (RFE) was employed. RFE iteratively removes the least important feature based on its ranking in the previous iteration. This process continues until the desired number of features was reached. The number of features is a balance between model accuracy, computational effort, the risk of overfitting and improved interpretability.

For feature selection, we employed a RF model with n_estimators set to 50. This number was chosen based on preliminary evaluations, which demonstrated that 50 estimators effectively allowed us to obtain stable feature rankings without incurring excessive computational costs. By focusing on the most relevant features, RFE helps improve model performance and reduces the risk of overfitting by preventing the model from relying on dataset specific information.

To compare the robustness of using an estimator-based feature selection method (Random Forest RFE) an all-relevant feature selection method (BorutaPy) was also tested. The preprocessing of categorial features using one hot encoding splits a single categorical column into a set of multiple binary columns so that they can be numerically manipulated. For Boruta feature selection, one hot encoding was replaced with an ordinal encoder in preprocessing. The impact of this change was negligible (< 0.003) for the XGB and Random Forest models for PRC-AUC performance. The Boruta feature selection was applied to the filtered dataset for all assays and the full feature set of 157

features. Two estimators, Random Forest(RF) and XGBoost(XGB), were used to eliminate unnecessary features. A variable number of features were selected by the Boruta package depending on the assay and the estimator used. The initial 157 possible features were pruned into the following sets:

Assay	Estimator	Feature count
Myeloid	RF	71
Haem	RF	116
TSO500	RF	51
Myeloid	XGB	17
Haem	XGB	28
TSO500	XGB	33

Inspection of the selected sets showed a limited intersection between them. It was also noted that the categorical variables corelated poorly with the RFE feature set compared to numerical features. This was likely the result of using ordinal encoding of categorical variables where there was no natural ordering of the classes e.g. genes.

To test a sample of the Boruta selected features, two sets, (Haem, XGB,28), (TSO500,XGB,33) were used build all models and apply them to the filtered dataset across all model estimators and assays. Comparing the RFE feature set with this sample of BorutaPy selected features showed an average decrease in the PRC_AUC metric of 0.34 for the TSO500 dataset and a decrease of 0.55 for the Haem dataset derived features. These results are shown in the Table S2 Boruta Feature Selection.xlsx.

On balance, the RFE derived feature set (35 features) performed better over a wider range of the datasets and was used in subsequent analysis.

Model evaluation

To create training and test sets for evaluating model performance, we created a custom variant of the Scikit learn TimeSeriesSplit cross validator [21]. Batched sequencing run analysis is temporal in nature and uses data from the expert curated variants of earlier runs. It is critical to only validate the models using sets of variants from sequencing runs occurring after the runs used in the training set. K cross validation folds were created by partitioning the variants by splitting runs into K + 1 sets. The Nth cross validation fold uses the earliest N partitions as the training set and the $N + 1$ partition as the test set. Overall, all data is used for both training and testing, preventing data leakage, and ensuring a realistic model evaluation. This approach recapitulates the time-based process where genomic analysis reuses the analysis of previous sequencing run analysis.

Model performance is often measured with F1 scores or Receiver Operating Characteristics (ROC) plots, but for strongly imbalanced dataset as frequently found in health studies ROC plots are less useful. Here the number of positive results, (reportable variants), are greatly outnumbered by the negative results (benign or technical artefact variants), so precision recall curves (PRC) and their associated area under the curve (AUC) were used for model performance metrics [22]. All performance metrics are found in Table S4 Model Performance Metrics.

Results

The tree ensemble estimator models (see Table 3), performed consistently well across the assays achieving between 0.833 and 0.997 for XGBoost and between 0.904 and 0.995 for Random Forest on the precision recall area under the curve (PRC AUC) metric. Both the LR model and the Neural Network model performed less consistently and poorly on the Myeloid assay with its smaller datasets. The test sets were chosen as described in the Model Evaluation section. It can also be seen in Table 3 that the full features set of 157 variables performed better than the RFE pruned features for all assays on the filtered dataset using either the Random Forest or the XGBoost estimator models.

The performance profiles of the PRC AUC data for the assays on the filtered dataset and the pruned feature set are shown in Fig. 1.

Time series cross validation was performed as described in the Model Evaluation section. Fig. 2 shows the performance for the XGBoost estimator. As the training set size increases with each CV fold, the PRC AUC can be seen to plateau, particularly for the TSO500 solid tumour assay, whose performance is below that of the haematological assays. This performance difference is likely due to the mean number of reportable raw variants relative to reportable variants per patient: TSO500 (total = 453.8, reportable = 2.9 (0.6%)), versus the Haem (total = 76.4, reportable = 2.0 (2.6%)) and Myeloid (total = 40.3, reportable = 0.8 (2.0%)) assays.

In addition, the percentage of novel variants for the TSO500 assay is approximately double the percentage for that of the Haem and Myeloid assays (see Fig. 3). The number of novel variants progressively decreases as the in-house knowledgebase accumulates data on more reportable variants over time. Up until June 2022, the Myeloid panel had 8 reportable genes, this increased to 22 reportable genes in July 2022 leading to the early variability of the Myeloid data points in Fig. 3. Both the Haem and the Myeloid

Table 3 Model performance on approximately 20% hold out test set. Best performance per group shown in bold. Additional performance metrics of precision, recall, F1 and ROC-AUC May be found in Table S4 model performance metrics

		PRC-AUC Performance		
Full dataset	**Features**	**Haem Assay**	**Myeloid Assay**	**TSO500 Assay**
Logistic Regression	35	0.909	0.899	0.852
Random Forest	35	**0.989**	0.973	**0.904**
XGBoost	35	**0.989**	**0.983**	0.833
Neural Network	35	0.984	0.916	0.882
Full data set size		444,073	101,871	804,074
Training set size (variants)		355,726	77,820	588,767
Test set size (variants)		88,347	24,051	215,307
Filtered dataset with Pruned Feature set				
Logistic Regression	35	0.982	0.979	0.887
Random Forest	35	0.994	0.985	**0.913**
XGBoost	35	**0.996**	**0.991**	0.900
Neural Network	35	0.993	0.980	0.891
Filtered dataset with Full Feature set				
Logistic Regression	157	0.953	0.955	0.890
Random Forest	157	0.995	0.991	**0.931**
XGBoost	157	**0.997**	**0.995**	0.921
Neural Network	157	0.817	0.824	0.873
Filtered data set size		65,985	15,208	29,979
Training set size (variants)		50,383	10,569	22,312
Test set size (variants)		15,194	4,541	7,506

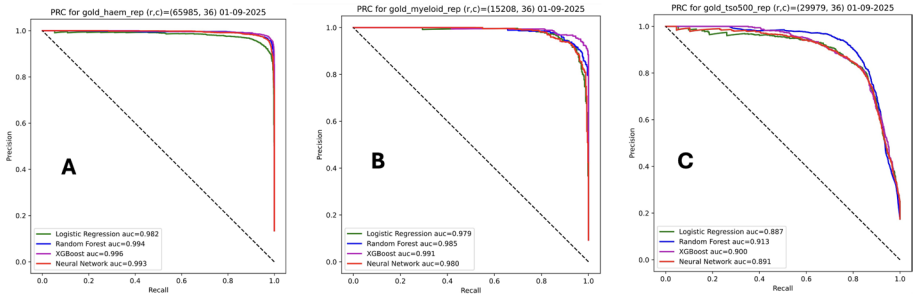

Fig. 1 Model estimators performance For filtered Haem dataset. **A** Precision Recall Curves (PRC) for the Haem assay and all model estimators. The filtered dataset for the Haem assay is referred to internally as "gold_haem_rep". Performance is on a hold out test set ($n = 15{,}194$) and model training set ($n = 50{,}383$). **B** Precision Recall Curves (PRC) for the Myeloid assay and all model estimators. Performance is on a hold out test set ($n = 4{,}541$) and model training set ($n = 10{,}569$). **C** Precision Recall Curves (PRC) for the TSO500 assay and all model estimators. Performance is on a hold out test set ($n = 7{,}506$) and model training set ($n = 22{,}312$)

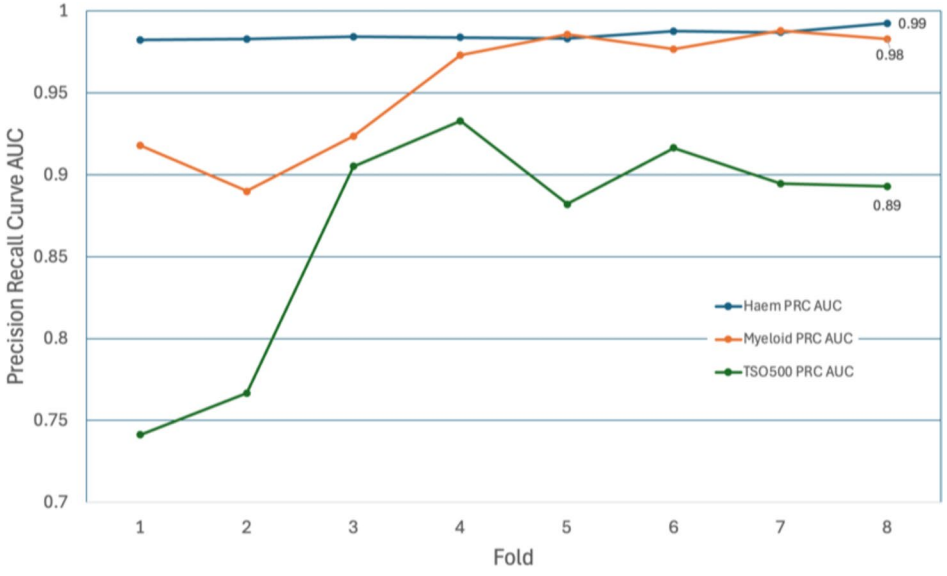

Fig. 2 Assay time series 8-fold cross validation performance. Precision recall curves AUC values for each assay for the XGBoost estimator

assay graphs start in October 2021 but benefit from reported variants curated prior to this date during assay validation and therefore have a lower number of novel variants initially.

The importance of features used in the models were calculated as discussed in the Data Extraction section. The relative importance for the Haem assay is shown in Fig. 4. The pruned feature names, their source and description are shown in Table 2. The features have been placed into four groups as follows; Sequencing and assay specific features (Seq), curated knowledge bases, both public and in-house (KB), reference genome variant effect (G) and public in-silico predictors (IS). The variability of importance across all the assays is shown in Figure S1. The feature importance was also averaged across all the assays and shown in Figure S2.

Each assay has a diverse range of genes, disease contexts and wet lab processes. Additionally, the bioinformatic and annotation pipelines vary greatly between the blood cancers (Haem and Myeloid) and the solid tumour cancers (TSO500). These factors suggest

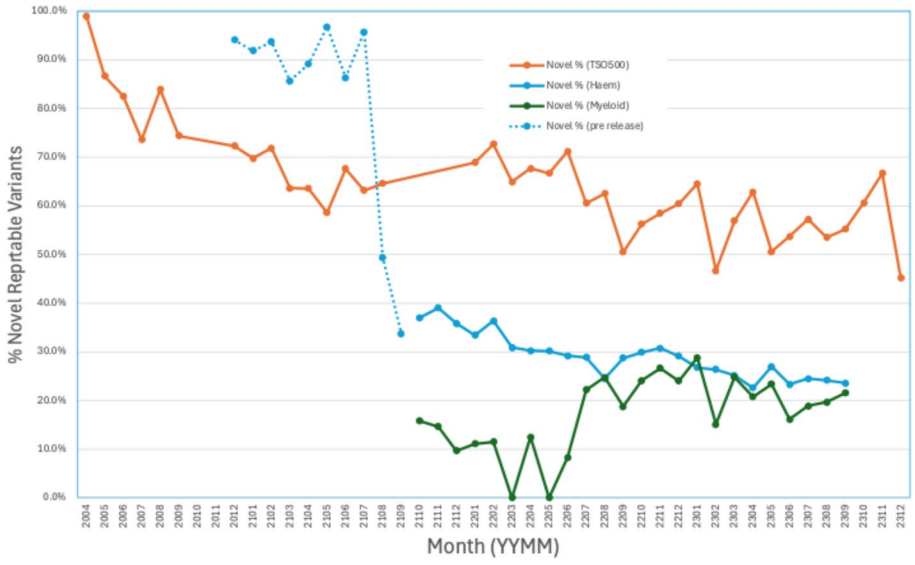

Fig. 3 Percentage of novel variants by assay and by month. The percent of novel reported variants by assay (TSO500: orange, Haem: blue and Myeloid: green) and month (yymm). Over time the number of novel variants reported generally decreases as the in-house database accumulates reported variants observed in patients assayed. The data for the Haem assay is shown as blue dotted during the validation phase of the assay

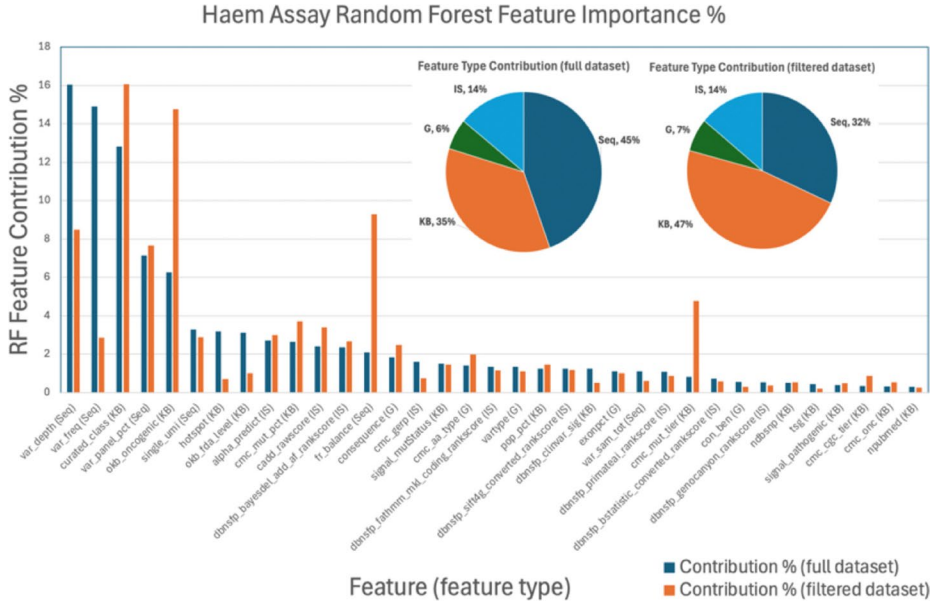

Fig. 4 Haem assay feature importance. Feature importance (%) for the Random Forest model for the full Haem Assay dataset (blue bars) and filtered dataset (orange bars). Features are after feature pruning. Features are annotated by feature type in parenthesis; sequencing and assay specific features (Seq), curated knowledge bases, both public and in-house (KB), reference genome variant effect (G) and public in-silico predictors (IS). The inset pie charts show the relative contribution of feature types for both the full and filtered datasets

a per assay feature set be used to build the models for each assay, an area for further research.

Model inference of variants from new sequencing runs can provide decision support for genetic scientists interpreting patient samples. For our clinical model implementation of individual variant predictions, we selected non-linear tree-based models that provide a quantitative interpretation of the contribution each model feature adds to the

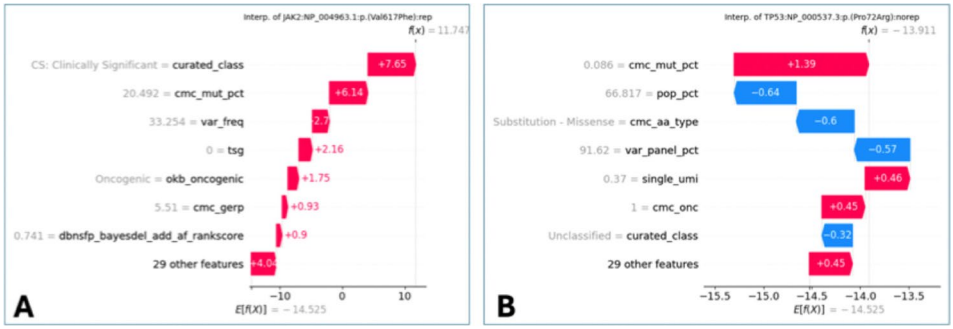

Fig. 5 Variant interpretation waterfall plots. Two example variant interpretation plots from the Haem assay. The variant on the left was reported and the right variant was not. **A** Interpretation plot for variant JAK2 (Val617Phe) predicting it should be clinically reported by the model with high scoring features of curated class and Cosmic Cancer Mutation Census. Features contributing to this prediction are shown in order of weight of contribution from largest (top) downwards. The annotation values of each feature are shown in grey to the left of the feature name. Red indicates a feature that positively contributes to a variant being reported while blue indicates a negative contribution. **B** Waterfall plot for the variant TP53 (Pro72Arg) showing its high prevalence in the general population (pop_pct = 66.8%) and its frequent occurrence in the panel samples (var_panel_pct = 91.6%) give a low prediction of it being reportable

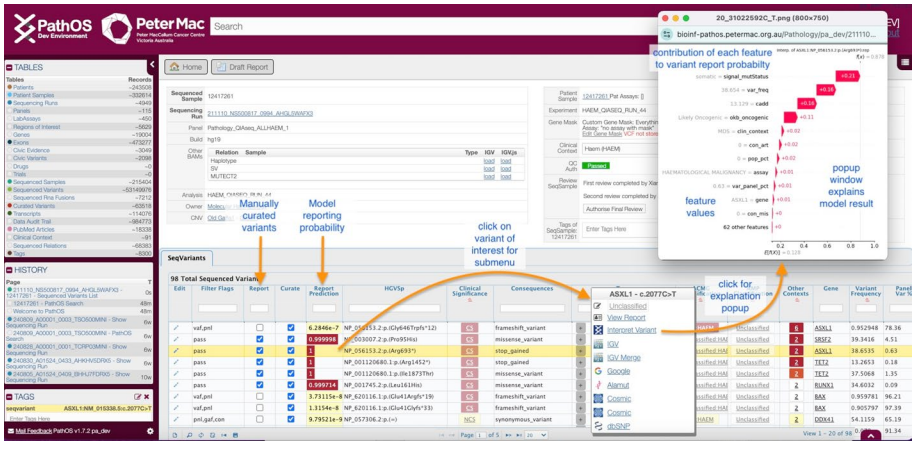

Fig. 6 Model integration with tertiary analysis platform (PathOS)

final prediction for each variant. Tree-based models such as Random Forests [23] and XGBoost [24] lend themselves to detailed interpretations of predictions using Shapley values, an approach from cooperative game theory [25]. The interpretation of the additive probability contributions of features can be most naturally represented as a waterfall plot where the features are sorted from top to bottom in order of decreasing contribution effect. See Fig. 5.

The interpretation plots are made available to the genomic scientist performing variant analysis through the in-house tertiary analysis platform (PathOS [3]). This allows patient variants to be sorted by model prediction likelihood and the viewing of individual variant model interpretation through a popup window displaying how the model used variants attributes (see Fig. 6). All genomic scientists and pathologists using PathOS are trained in variant interpretation and have knowledge of the model features. We are in the process of actively soliciting user feedback on this integration and will incorporate user feedback. We anticipate the richer quantitative variant data that the model provides to users will allow them to better discriminate more difficult cases.

Discussion

This study created machine learning models to assess suitability for providing decision support for clinical reporting of NGS targeted assays. The performance of tree-based ensemble models (Random Forest and XGBoost) on all assays ranged from 0.904 to 0.997 PRC AUC. The mean PRC AUC results of tree-based ensemble models show progressive improvement down Table 3: (full dataset with pruned features($n = 35$)) mean = 0.945; (filtered dataset with pruned features($n = 35$)) mean = 0.963; (filtered dataset with full features($n = 157$)) mean = 0.972.

Using all features ($n = 157$) rather than the pruned features ($n = 35$) from recursive feature elimination selection proved superior for this study. Despite missing data, redundancies and correlated features, using the full set of features proved an advantage although this improvement comes at a computational cost as well as the risk of overfitting. The sourcing of comparable datasets from other institutes is a top priority for future study into the generalizability of in-house and external annotation features in diagnostic oncology.

There was an incremental improvement when using the filtered data that used cutoff values to remove variants unlikely to be diagnostically relevant for reporting. Filtering variants greatly reduced the volume of unreported variants while maintaining reportable variants and allowing the models to focus on identifying reportable variants rather than excluding unreported variants. While providing a slight improvement in performance, filtering with hard cutoff criteria risks excluding variants that may be reportable near the limit of cutoff values. For both the full and filtered datasets, the solid tumour (Illumina TSO500) assay did not perform as well as the haematological assays (Fig. 2; Table 3) although the model performances (PRC AUC of full dataset 0.904 and filtered dataset 0.913) are still likely to provide a significant benefit for genomic analysts.

Comparison of which features drove model performance showed clear differences between models built on the full datasets compared to the filtered variant data. Features derived from the sequencing process (variant depth, variant frequency etc.), are the most important feature type (Seq = 45%) compared with knowledge-based features (KB = 35%) for the full Haem dataset. In contrast, for the filtered dataset, the situation is reversed (Seq = 32%, KB = 47%). This can be understood by noting the sequencing-based features are used to filter out technical artefacts from the NGS process that are dominant in the full datasets. In the filtered dataset, where many of these variants are removed, the model more fully utilises features from variant knowledgebases to identify reportable variants. For the Haem assay filtered dataset, the knowledge base features of most importance were: curated_class 16.1% (in-house curation DB), okb_oncogenic 14.8% MSK OncoKB and cmc_mut_tier 4.8% Cosmic Cancer Mutation Census. For the TSO500 assay filtered dataset, the two top features where okb_oncogenic 15.4% and signal_mutstatus 9.0% MSK Signal DB. It is noteworthy that of these data sources, Memorial Sloan Kettering's Signal database is the only one available for non-research use. The usage restrictions on large curated variant databases are an emerging issue for clinical laboratories wishing to improve their curation processes. The large contribution of in-house assay sequencing features (Seq) to model performance (full dataset: 45%, filtered dataset: 32%) represent features that are difficult to be generalised to other assays. In addition, these features may not generalise easily between laboratories running the same assay due to differences in SOPs, sample preparation and local conditions.

The dependency on sequencing-based features for all models built from the full dataset or filtered datasets suggests that generalising these types of models between laboratories or assays will be difficult. The models depend on features and statistics derived from the actual sequencing process (feature group Seq) which are not present public variant databases (see Fig. 4).

A key weakness of many initiatives to introduce AI into healthcare stem from the adoption of opaque or 'black box' models which don't make transparent the interpretation of individual predictions. This issue has been raised multiple times in the literature [26, 27], and it is becoming less acceptable for models in healthcare to only predict outcomes without explanation. Aggregated model feature importance is commonly available but local interpretation of individual predictions is more difficult. Understanding individual predictions is necessary for trust, accountability and to confidently action model results. Individual prediction interpretations have been classed as a *mandatory criteria* in the recently released AI in Healthcare Guidelines released by the Dutch Ministry of Public Health [28, 29]. This has also been driven by the European GDPR and other regulations which further encourage a 'right to explanation' for AI systems [30]. Individual variation interpretations can also reduce Automation Bias (AB), the tendency of over relying on automation [31].

Conclusion

The implementation of machine learning models in the clinical reporting of NGS assays is emerging as AI is applied to variant interpretation in oncology. This study has shown data derived from targeted NGS assays can serve as a robust base for generating predictive models that support genetic scientists in clinical decision-making. We have shown machine learning models can achieve validation PRC AUC scores ranging from 0.997 to 0.904 across targeted hematological and solid tumour assays. The ability to classify clinically reportable variants provides decision support to manual expert curation, a significant bottleneck in scaling up clinical sequencing operations. Machine learning models allow the analysis of large numbers of variant annotations that can be complementary, redundant and sometimes conflicting. It allows for a more thorough interpretation of a variant's consequence within a patient's clinical context, a task that is demanding of trained genetic scientists. These models' provision of explanatory waterfall plots for individual variant's prediction further aids scientists in evaluating and understanding the contributing factors to each decision, resulting in confidence in the decision support algorithms and the reporting process. The integration within the tertiary analysis workflow provides the much needed transparency of AI into critical healthcare systems. The integration also allows for future work to measure clinical efficiencies of implementing ML models and collect valuable feedback of scientists working on variant analysis. Another area for future study is the generalisation of the models across multiple assays and between laboratories to broaden the benefits of machine learning. The application of AI into the clinical reporting workflows has just begun and will allow the streamlining the reporting process and reduce inter-reviewer variability to allow for more accurate, robust and clinically defensible oncology reporting.

Supplementary methods

For information regarding the diagnostic service please see https://www.petermac.org/molecular-pathology.

Haem assay and myeloid assay – QIAseq custom design

Targeted sequencing of 57 genes was performed using custom QIAGEN QIAseq single primer extension-based target enrichment and Illumina NextSeq500 with 150 bp paired end read sequencing. A customised CLC Genomics Workbench (QIAGEN) analysis pipeline (v3.2) was used to generate aligned reads and call variants (single nucleotide variants and short insertions or deletions) against the hg19 human reference genome. Variants were analysed using PathOS software (Peter Mac) and described according to HGVS nomenclature. The 'Haem' assay included assessment of the full gene panel ($n = 57$ genes) and the 'Myeloid' panel included assessment of a limited set of either eight genes (between October 2021 and June 2022) or 22 genes (between July 2022 and December 2023).

Solid tumour assay - TSO500

Targeted sequencing of 523 cancer genes from DNA and 55 cancer genes from RNA was performed using the Illumina TruSightTM Oncology 500 assay (TSO500). DNA and RNA were extracted from pathologist-selected areas of submitted formalin-fixed paraffin embedded (FFPE) tumour using the Qiagen AllPrep DNA/RNA FFPE kit. RNA was reverse transcribed to cDNA. DNA was ultra-sonically sheared to an appropriate size range. A 'Unique Molecular Identifier' (UMI) tagged DNA library and a standard RNA library were prepared and enriched with magnetic streptavidin beads following targeted hybridisation to gene-specific biotinylated probes. Pooled, normalised libraries were sequenced to an appropriate target mean coverage on an Illumina sequencing platform (NextSeq or NovaSeq). Illumina Software TSO500 v2.2 Local App was used to generate aligned reads and call variants against the hg37 human reference genome. Clinical Genomics Workspace (CGW) from PierianDx and Navify Mutation Profiler from Roche were used to annotate, filter and report clinically relevant findings.

Abbreviations

API	Application Programming Interface
AIPA	AI Prediction Algorithm
AUC	Area Under the Curve
BAM	Binary Alignment Map format
CADD	Combined Annotation Dependent Depletion
cDNA	complementary DNA
CDSS	Clinical Decision Support System
delins	A variant which combines a deletion and an insertion
FDA	U.S. Food and Drug Administration
FN	False Negatives
FP	False Positives
HGVS	Human Genome Variant Society
indel	Insertion/Deletion
MNP	Multi-Nucleotide Polymorphism
MRD	Measurable Residual Disease
NGS	Next Generation Sequencing
NATA	National Association of Testing Authorities
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PRC	Precision Recall Curve
ROC	Receiver Operating Characteristics
TP	True Positives
TSV	Tab Separated Variable format

UMI Unique Molecular Identifier
VAF Variant Allele Frequency
VCF Variant Call Format

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00489-y.

Additional File 1: Figure S1.pdf

Additional File 2: Figure S2.pdf

Additional File 3: Figure S3 Workflow.pptx

Additional File 4: Table S1 All Features Correlation Matrix.xlsx

Additional File 5: Table S2 Boruta Feature Selection.xlsx

Additional File 6: Table S3 Disease Count.xlsx

Additional File 7: Table S4 Model Performance Metrics.xlsx

Additional File 8: Table S5 Model Training Times.xlsx

Additional File 9: Table S6 SNP Count.xlsx

Acknowledgements
The authors would like to acknowledge the generosity of our funders in making this project possible.

Authors' contributions
K.D. conceived the study. K.D. and R.P. wrote the software and wrote the manuscript. Ongoing feedback and advice given by A.F., Y.K., R.L., E.T.,P.B. and S.F. All authors read and approved the final manuscript.

Funding
This research was supported by the Laby Foundation, The Peter Mac Foundation, Therapeutics Innovation Australia and a National Health and Medical Research Council (NHMRC) Program Grant (1054618). The research benefitted by support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support.

Data availability
The data used for this research is clinical patient data covered by ethics approval (HREC: 81837) which precludes making data generally available. Please contact the authors regarding potential collaboration on this work.

Declarations

Ethics approval and consent to participate
Ethics approval was granted on 1 April 2022 (EPIC Study Code: PMC81837, HREC: 81837).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 June 2025 / Accepted: 29 September 2025
Published online: 29 October 2025

References

1. Doig K, Papenfuss AT, Fox S. Clinical cancer genomic analysis: data engineering required. Lancet Oncol. 2015;16:1015–7. https://doi.org/10.1016/S1470-2045(15)00195-3.
2. Doig KD, et al. Findings from precision oncology in the clinic: rare, novel variants are a significant contributor to scaling molecular diagnostics. BMC Med Genomics. 2022. https://doi.org/10.1186/s12920-022-01214-y.
3. Doig KD, et al. PathOS: a decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories. Genome Med. 2017;9:38. https://doi.org/10.1186/s13073-017-0427-z.
4. Hoskinson DC, Dubuc AM, Mason-Suares H. The current state of clinical interpretation of sequence variants. Curr Opin Genet Dev. 2017;42:33–9. https://doi.org/10.1016/j.gde.2017.01.001.
5. Lincoln SE, et al. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genet Med. 2021;23:1673–80. https://doi.org/10.1038/s41436-021-01187-w.
6. Ainscough BJ, et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nat Genet. 2018;50:1735–43. https://doi.org/10.1038/s41588-018-0257-y.
7. Wagner AH, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet. 2020;52:448–57. https://doi.org/10.1038/s41588-020-0603-8.

8. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. Genet Med. 2015;17:405–24. https://doi.org/10.1038/gim.2015.30.
9. Li MM, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the association for molecular Pathology, American society of clinical Oncology, and college of American pathologists. J Mol Diagnostics: JMD. 2017;19:4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002.
10. Tate JG, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2018;47:D941–7. https://doi.org/10.1093/nar/gky1015.
11. Cheng J, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. Science. 2023;381:eadg7492. https://doi.org/10.1126/science.adg7492.
12. Chang MT, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34:155–63. https://doi.org/10.1038/nbt.3391.
13. Chang MT, et al. Accelerating discovery of functional mutant alleles in cancer. Cancer Discov. 2018;8:174–83. https://doi.org/10.1158/2159-8290.Cd-17-0321.
14. Liu X, Li C, Mou C, Dong Y, Tu Y. DbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103. https://doi.org/10.1186/s13073-020-00803-9.
15. Schubach M, Maass T, Nazaretyan L, Roner S, Kircher M. CADD v1.7: using protein Language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res. 2024;52:D1143–54. https://doi.org/10.1093/nar/gkad989.
16. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4. https://doi.org/10.1038/ng.3774.
17. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2015. https://doi.org/10.1093/nar/gkv1222.
18. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. https://doi.org/10.1038/s41586-020-2308-7.
19. Chakravarty D, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017. https://doi.org/10.1200/po.17.00011 .
20. Srinivasan P, et al. The context-specific role of germline pathogenicity in tumorigenesis. Nat Genet. 2021;53:1577–85. https://doi.org/10.1038/s41588-021-00949-1.
21. TimeSeriesSplit. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.
22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.
23. Ho TK. Random Decision Forests. Proceedings of the Third International Conference onDocument Analysis and Recognition. 1995;1:278–82. https://doi.org/10.1109/ICDAR.1995.598994.
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery large-scale machine learning, SanFrancisco, California, USA. https://doi.org/10.1145/2939672.2939785.
25. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
26. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. https://doi.org/10.1186/s12916-019-1426-2.
27. Amann J, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20:310. https://doi.org/10.1186/s12911-020-01332-6.
28. Rakers MM, van Buchem MM, Kucenko S, et al. Availability of evidence forpredictive machine learning algorithms in primary care: a systematic review. JAMA Netw Open. 2024;7(9):e2432990. https://doi.org/10.1001/jamanetworkopen.2024.32990.
29. de Hond AAH, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022. https://doi.org/10.1038/s41746-021-00549-7.
30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x.
31. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19:121–7. https://doi.org/10.1136/amiajnl-2011-000089.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.