Predicting and Avoiding Dog Barking Behaviour through Deep Learning

Zoe Li, Richard O. Sinnott
School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia, Contact: rsinnott@unimelb.edu.au

ABSTRACT

Excessive and intense barking can be a problem for dog owners and households more generally. Dogs can react to (be triggered by) a wide range of environmental noises and situations, e.g., the sound of the postman, the doorbell amongst a whole range of other sounds. In many situations this behaviour results in reprimands and/or punishments, e.g., demands for the dog to be quiet. Veterinary scientists and dog behaviour experts increasingly recognise that dogs respond more positively to positive reinforcement. Thus, instead of shouting at the dog, if the sounds can result in treats being given, then it is possible to change the dog's behaviour, i.e., the dog recognises the sound as a positive and not a negative thing hence it does not bark. Modifying a dog's behaviour when the owner is present is possible, but many pets are left alone for periods of time hence a solution is needed to automate the positive reinforcement. This is the focus of this paper. We demonstrate how rich and diverse urban audio datasets can be used as the basis for automatically capturing Mel-Frequency Cepstral Coefficients and Mel Spectrogram features. These are then used in conjunction with Convolutional Neural Networks to learn and hence identify and predict environmental sounds that precede a dog's barking. We achieve an overall accuracy of up to 87.6% based on 50 representative environmental sound classes.

KEYWORDS

Keywords, Audio recognition, Convolutional Neural Networks, Mel Frequency Cepstral Coefficients, Dog Barking, Behaviour Prediction

ACM Reference Format:
Zoe Li, Richard O. Sinnott. 2024. Predicting and Avoiding Dog Barking Behaviour through Deep Learning. In *2024 Australasian Computer Science Week (ACSW 2024), January 29–February 02, 2024, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3641142.3641176

1 INTRODUCTION

Sound recognition has also been explored in many fields, including automatic speech recognition and environmental sound classification [10], as well as in many situational contexts including smart homes and security monitoring. The emergence of audio recognition based on the development of deep learning machine and neural networks has now made it possible to tackle complex audio classification tasks. Convolutional neural networks (CNN) and their ability to learn patterns and then to identify and distinguish differences in data has made it possible to address many diverse research topics in the sound classification field [26].

In this paper we explore a novel application scenario: the potential cause of dog barking based on environmental sounds. Due to the large variety and number of sounds in urban settings, it is the case that many of these can cause canine anxiety resulting in excessive barking. Dog owners seeking to train their pets to minimize such behaviours have little recourse, especially if they are not always in the household with the pet at the given time when the barking occurs. Such barking may however, impact on any neighbours who may be affected by the sound of the dog barking when no-one is home.

This prediction of dog barking has many challenges: the large amount and diversity of urban noises, e.g., a doorbell can have many different forms; the difficulty in collecting a large amount of data associated with each potential barking trigger as well as the individual challenges based on the breed, characteristic and environment of each individual dog. Thus, some dogs may be triggered by footsteps near to the house, a doorbell, others by the sound of the refuge collection lorries, others by the sound of motorcycles etc.

This paper explores a machine learning-based approach to automate the prediction of barking behaviour of dogs. The ultimate goal is to develop an Internet-of-Things (IoT) device that release treats/snacks when a specific noise or collection of noises is identified to associate any triggering sounds to encourage positive behaviour of dogs.

The rest of this paper is structured as follows. In Section 2 we present a literature review focused upon audio recognition, environmental sound classification systems, and existing datasets. In Section 3 we introduce the methodology used in the research, comparing different data processing, feature extraction, data augmentation techniques. In Section 4 we describe the experimental set up and preparing the data sets. In Section 5 we present the experiment results with dog barking classification. Finally in Section 6 we conclude the work and identify the limitations and potential future research directions.

2 RELATED WORKS

The earliest research into voice recognition focused on automatic speech recognition (ASR). Due to limitations in computing power as well as a lack of sound analysis technology, the primary approach was pattern matching based on acoustic principles [6]. The first big breakthrough in sound recognition utilised Hidden Markov Models (HMMs). The statistical power of HMM makes them suitable for

ACSW 2024, January 29–February 02, 2024, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1730-7/24/01
https://doi.org/10.1145/3641142.3641176

a wide range of sound classification tasks. With the exponential increase in computer computing power and widespread availability of audio files, an explosion in automatic speech recognition including large vocabulary continuous speech recognition systems has now taken place. Due to advancements in deep learning, neural networks have demonstrated exceptional suitability for recognition tasks combined for example with Support Vector Machines (SVMs) [14]. CNNs are now the predominant approach for voice recognition tasks [22].

Classification of environmental noise and background sound more generally has benefited from improvements in feature extraction from audio files. This includes time features, frequency features, and time-frequency features such as Mel Frequency Cepstral Coefficients (MFCC) [4]. Environmental sound classification is now widely used in many home automation scenarios [25] as well as species classification tasks, e.g., bird call classification [3].

There has been extensive work on binary classification of sounds. Numerous studies have explored biological characteristics classification using random forest trees [21] and logistic regression [20]. The number of studies using machine learning to predict behaviour based on diverse sounds is rather more limited [13].

Research on dog behaviour has identified that dog barking can be caused by many factors: dogs may bark when they are fearful, due to noise and/or separation anxiety. Noise from the surrounding environment have been shown to be a trigger to dogs and their barking behaviour [24]. Indeed, many dogs are bred specifically to have this trait, e.g., guard dogs have sensitive hearing and are expected to bark to warn of potential intruders.

Dog barking also plays a role in communication. Dog barking is clearly different between dogs, e.g., a Chihuahua bark is different to a Great Dane bark [16]. Ideally a dog's real-time barking behaviour should factor in the surroundings to the model to analyse the cause of the specific dog barking from multiple perspectives to improve the accuracy of prediction, i.e., there may not be a single cause of the barking, but several factors and sounds that give rise to the barking behaviour.

More and more dog owners and trainers focus on positive reinforcement to encourage good behaviour compared to negative punishment. The use of treats to reward dogs was shown to be positively correlated with dog obedience in over 50% of households [9]. This suggests that intervening and preventing barking with treats, can calm a dog's anxiety to different triggering sounds.

There are many large-scale environmental sound datasets that exist. These datasets include different types of environmental sounds such as city streets, forests, traffic noise, etc. Two databases widely used for urban environmental sound classification tasks are UrbanSound8K [18] and ESC-50 [15]. Two databases consist of a collection of urban sounds. These include 8,732 common sounds heard in a city including dog barking. Waveform signals need to be extracted from the database sound files for feature extraction.

3 DATA PRE-PROCESSING AND MODEL STRUCTURE

The high-level approach taken in this work includes a multi-class classifier inspired by [22] based on a CNN-based model. The emphasis of this model is on using appropriate convolution kernels to efficiently extract waveform signal features. To achieve the best training results for this classification task, various CNN structures, data inputs, pooling layers, and learning rate parameters were explored.

As a model for detecting environmental sound events, it is necessary to classify a wide range of different environmental sounds. The dataset used for the first ESC classification task was the ESC-50. This has 50 types of common urban sounds based on several higher-level sound categories: animal sounds, natural soundscapes, human non-speech sounds, interior/domestic sounds, and exterior/urban noises. These categories include a range of possible sounds that a domestic dog may hear and trigger barking in their home environment. The dataset contains a CSV file and over 2000 .wav format audio clips. We use a Python library (LibroSA) to read the .wav files and include them in a one-dimensional array based on the time series representation of the audio sample rate.

In the ESC-50 dataset, the 2000 audio files each category containing 40 samples of different sounds. Each of these audio files has a duration of 5 seconds. This ensures an evenly distributed representation across all sound categories hence no further data balancing or sampling process is required. For each audio file, we normalise the waveform so that the volume level is consistent and apply standard deviation to centre them to the zero means so that sound distortion can be avoided. This process also minimises the difference between audio files in the same category caused by different environmental conditions and any external differences in acoustic characteristics between the different classes.

Table 1 shows the different audio training input lengths used for the experiments. Using a base CNN model, we use the original 5-second-long audio files; normalized 5-second-long audio files; normalized random 2-second-long audio files; normalized random 2-second-long audio files, selected non-silence intervals; normalized random 3-second-long audio files, normalized random 3-second-long audio files, and selected non-silence intervals using extracts from the basic MFCC features as input. The normalized 5-second-long audio is used as the base audio input as this was shown to have the highest degree of accuracy.

To support non-silent intervals, we used the function *librosa.effects.split()* to split all non-silent segments, however this often captures small nonsensical segments that are meaningless for model training. For example, some human sounds have intermittent acoustic features, such as coughing and snoring and shorter durations reduce the model's ability to discriminate audio. Although this reduces the amount of input data and the training time, the results suggest that silent sections in the audio contribute positively to distinguishing sounds. Furthermore, balancing the dataset ensures that the model produces a fair and accurate prediction across all classes. Normalization of the data ensures that the volume levels of the audio files are consistent and prevent any outlier excessively impacting the model, whilst helping the model to capture differences between audio files more accurately. By adjusting the length of the input audio, we aim to find an optimal audio input length that can be selected for the model training and experiments.

MFCC and Mel Spectrograms are the dominant approaches used for extracting features from waveform signals. Mel Spectrograms provide a visual representation of the sound spectrum (see Figure

Table 1: Pre-processing methods and accuracy

Pre-processing method	Accuracy
5s Original Clip	0.63
5s Normalised	0.66
Random Normalised 2s	0.49
Most Non-Silent Interval Normalised 2s	0.46
Random Normalised 3s	0.52
Most Non-Silent Interval Normalised 3s	0.49

Table 2: 50 semantical classes in 5 major categories

Animals	**Natural soundscapes & water sounds**	**Human, non-speech sounds**	**Interior sounds**	**Exterior noises**
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Handsaw

1a), where the original audio signal is transformed into a spectrogram, which shifts audio information from the time domain to the frequency domain. With this approach the vertical axis represents frequency, and the horizontal axis represents time. The colour typically represents energy or amplitude.

MFCC is a standard technique used in audio recognition. It provides a set of eigen-coefficients extracted from the Mel Spectrogram that capture characteristics such as pitch, timbre, and properties related to the perception of sound. The MFCC extraction process includes pre-processing, use of Fast Fourier transforms, Mel scaling, logarithmic operations, discrete cosine transforms, dynamic feature extraction amongst other steps. It also includes smoothing of the spectrum by reinforcing the wave at low frequencies for human hearing range. Dogs have better hearing and a larger frequency range than humans [2], however the main target is to distinguish environmental sounds. MFCC's ability to compress wave feature dimensions makes it suited to classification tasks.

Figure 1 gives a visualised view of the Mel Spectrograms and MFCC of the 50 sounds selected as the basis for the work here. The Mel Spectrogram represents the time and frequency characteristics of the waveform. The MFCC coefficients (Figure 1 (right) give a fair representation of the acoustic characteristics of each class even though the dimensionality is reduced. MFCC conveys the waveform features well in the low-frequency area.

Table 2 enumerates the specific sound classes depicted in Figure 1. The 50 environmental sound features have been arranged into 5 different categories that covering possible urban sounds that might be found in each urban environment. The MFCC features show representative patterns illustrating that low-frequency features have been correctly interpreted in the pre-processing steps. For example, the top left graph of the Mel-spectrogram in Figure 1 (left) presents the example of a dog barking sound, since dogs are likely to respond to the sound of other dogs barking to defend their territory. The acoustic features of the dog bark are well represented in Figure 1a - a short sharp waveform shaped feature. The other sounds are shown in more detail in Table 2.

We trained the basic CNN model using both MFCC and Mel Spectrogram features. As a preliminary result, using the MFCC feature gave an accuracy of around 63% whilst using Mel Spectrogram gave an accuracy of 57%. Since we aim for training efficiency and accuracy, we combine both MFCC and Mel Spectrograms features to train the model. While MFCC gives a good representation, we need to keep higher frequency sounds as a feature since dogs are especially sensitive to higher pitch noises. Both features are extracted from the pre-processed video clip and saved in array format. To achieve this, we use an encoder to compress the dimensionality and size of the Mel Spectrogram features. Compressing the data size while keeping the features as much as possible is used to save time during the training process.

The baseline CNN used was based on a multi-input model with two branches. The MFCC features are processed on one path, and the Mel Spectrogram features processed on the other path. The input data was handled individually by each branch using convolutional layers, pooling, and fully connected layers. The pooling layers reduce computational complexity and feature dimensionality. The outputs of the two branches were then combined and

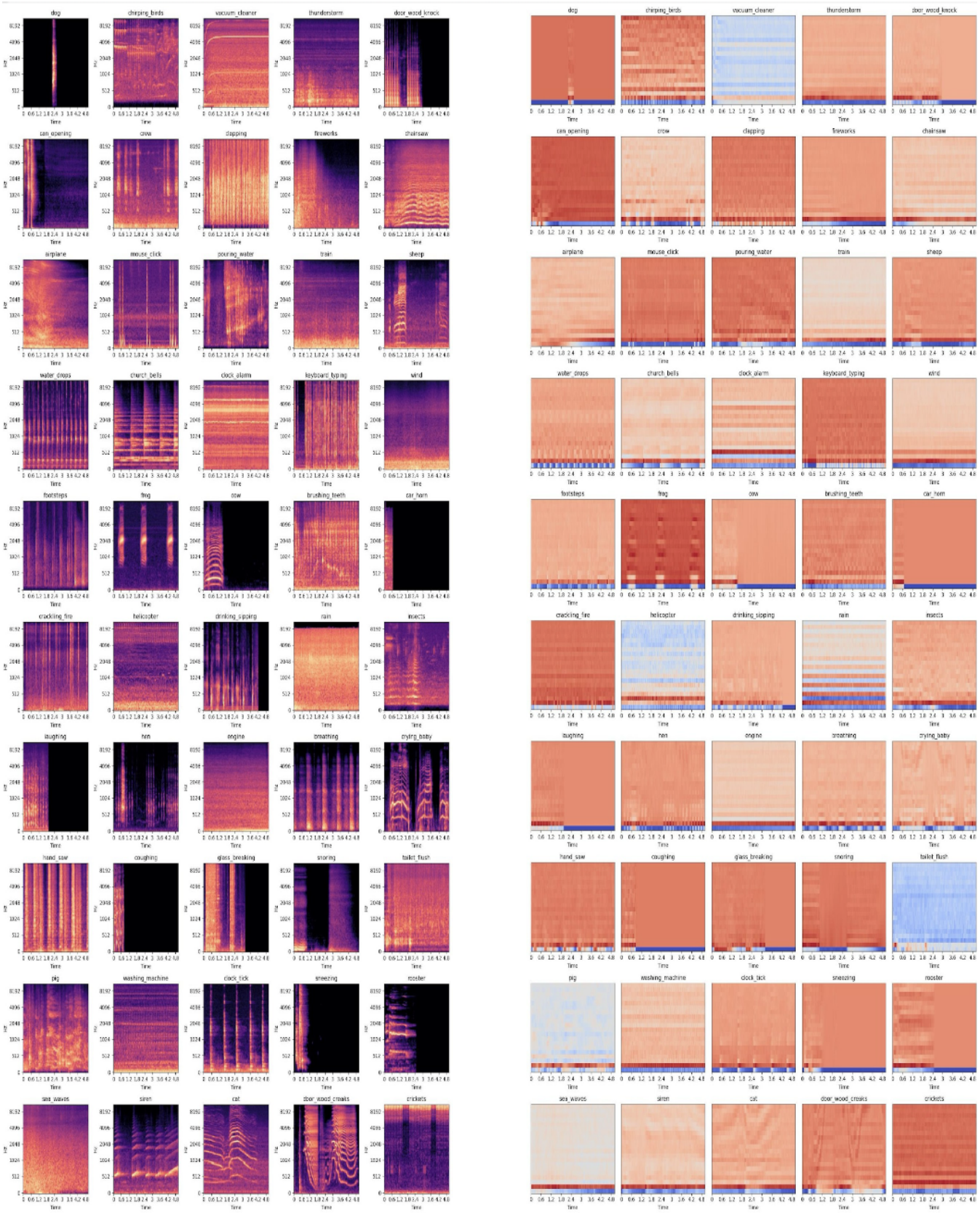

Figure 1: Features over samples of 50 audio sound classes

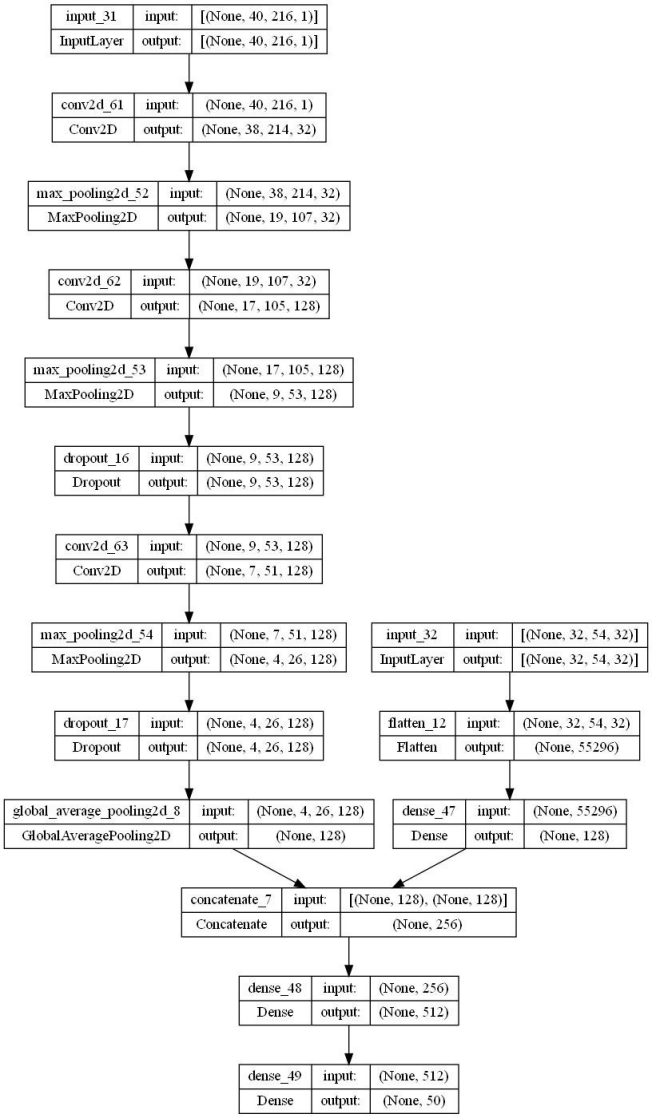

Figure 2: Multi-input CNN Structure

fed through a dense layer for final processing and prediction. The overall CNN model is shown in Figure 2.

The raw Mel Spectrogram feature had a dimension of (128, 216). To save training time, a *MinMaxScaler* was used to scale the Mel Spectrogram data to the [0, 1] range to reduce the dimension of the features. This also speeds up the time of the model to converge by minimising the interval of the features and preventing large values from impacting the model. A Rectified Linear Unit (ReLU) activation function was used to process the Mel Spectrogram input through two 2D convolutional layers comprising 16 and 32 filters, with max pooling used to lower the spatial dimensionality after each convolution layer.

The output was then flattened to connect to a Convolutional Autoencoder to further reduce the number of dimensions. The Convolutional Autoencoder utilised *Conv2D*, *MaxPooling2D* and *UpSampling2D* layers. A fully linked layer with an encoding dimension of 50 and a ReLU activation function in the encoder was used to process the output of the above steps. The encoded Mel Spectrogram data used a Dense layer to process the encoded Mel Spectrogram data to reduce the weight in the combined CNN model.

The MFCC path was used to process the MFCC features. These represent the main features used in the classifier. These are made up of *Conv2D* and *MaxPooling2D* layers, as well as a Dropout layer for regularisation and a *GlobalAveragePooling2D* layer for converting matrix data into a single vector.

A Concatenate layer was used to combine the output of the MFCC path with the Mel Spectrogram path and feed the output to a SoftMax activation function to execute the classification of environmental sounds.

A convolutional kernel of size (3, 3) was used. This size enabled efficient and detail-preserving waveform feature capture. The number of filters affects the complexity of the CNN learning of features. For the first convolutional layer, 32 filters were used to interpret the basic time and frequency features in MFCC, and then two layers comprising 128 filters was used to capture more detailed characteristics of the MFCC feature. The ReLU activation function was used here. This provides a computationally efficient way to save training time and avoid vanishing gradient problems [23].

To reduce the complexity of the features, the Max Pooling layer set the value of the pooling size range to the max value. The pooling layers were set to (2,2) to reduce the complexity to approximately half size to improve the training efficiency.

A dropout layer was used to avoid overfitting the model and improve its generalisability. The dropout rate was set to 0.3, i.e., 30% of the input unit was set to zero, effectively reducing interdependencies between the input units.

The number of neurons in the fully connected layers was set to 128 and 512 as deemed appropriate for the extraction of the MFCC features. The output layer had 50 neurons and used the SoftMax activation since there were 50 output classes. The SoftMax function transformed the output into a probability distribution.

The Adaptive Moment Estimation Algorithm (Adam) optimization algorithm was used for the models. Adam has an adaptive learning rate and supports efficient parameter updates by updating the weight between each layer of the neural network. The loss function was based on categorical cross-entropy. Cross-entropy loss of the predicted class probabilities and the true category labels was calculated in each batch. Categorical cross entropy was used to generate smooth and continuous gradients in the back-propagation process, thereby assisting in the reduction of loss when employing stochastic gradient descent (SGD) [19].

Considering the training efficiency, the batch size was set to 8 for rapid weight updates and faster convergence. The number of training epochs was set to 40. This was based on multiple training sessions to ensure enough training cycles for model accuracy whilst preventing overfitting.

The core classification task was based on a binary classification aimed at predicting whether a dog will bark or not. Training a model to predict the behaviour of a single dog poses challenges due to unique behaviour patterns and limited data availability for given dogs. Gathering sufficient data for each dog is time-consuming and impractical. Since each dog may behave differently, the training

data thus contained combined sounds from the UrbanSound8k data set. Separate noises and barks were combined into longer audio files using audio clips. Multiple models were then built and compared for this classification task with the aim of identifying the best-performing model.

4 MODEL SELECTION AND SCENARIOS

Numerous machine learning models exist that can form the basis for the sound classification preceding a given dog barking. In this paper we explore use of SVM, Random Forest, Logistic Regression, and CNNs. Data augmentation, regularization, and hyperparameter tuning techniques were also explored and the extent that they improve the model performance.

As noted, the data used to predict dog barking behaviour was based on the UrbanSound8K dataset using a model trained on the ESC-50 data set. UrbanSound8K provides an aggregation of 8,732 common (urban) sounds over 10 classes, including dog barking. The categories covered in the dataset include air conditioners, car horns, children playing, dogs barking, drilling, engine idling, gunshot, jackhammers, sirens, and background street music. All of these sounds are potentially likely to be heard in domestic settings where dogs are likely to be found.

As identified, it is clearly unrealistic to use the same model to match the unique barking behaviour of a specific dog, since each dog may well react differently to each sound. One approach to tackle this problem is to continuously collect sounds to prevent missing any possible trigger and classifying the sounds heard. After the dog barking sound is detected, the entire audio can then be used as input data. If the sound before the dog barking is directly followed by a barking sound, then we can consider this sound as a trigger for a given barking behaviour. However, there may be multiple sounds and combinations of sound that occur before a given dog barking behaviour occurs. To tackle this, we intersperse the dog barking audio files into other audio files to simulate realistic and complex sound scenarios associated with barking dogs. Furthermore, different dogs may respond to different sounds and sound combinations. To tackle this, we simulate multiple different dogs and the combinations of sounds that trigger their barking.

We thus utilise numerous different audio file combinations as the basis for classifying when a given dog might bark. These are listed in Table 2 below. We also assign these combinations of sounds to representative dogs that form the basis for the dog barking audio samples and the experiments that we perform.

We divide the barking behaviour into three main categories:

- dogs 1-3 have a single specific sound that always triggers (100%) their barking behaviour,
- dogs 4-6 would have two sounds that sometimes triggers (10-90%) their barking behaviour,
- dogs 7-10 have three different sounds that often triggers (30-80%) their barking behaviour.

We use the audio samples from the UrbanSound8K dataset to splice the sound files into a series of audio clips, with a 1-second break between each sound for training purposes. This allows us to use the environmental sound classification model to recognize the sound of a barking dog and use the prediction result as input for the following machine-learning training task. Figure 3 shows the typical representation of the sound of a barking dog interspersed with other environmental sounds with some sounds acting as a trigger for the dog barking.

Table 3: Dog barking behaviour patterns

Dog no.	Action
1	100% barks when 1 is heard
2	100% barks when 2 is heard
3	100% barks when 5 is heard
4	10% barks when 7 is heard, 90% barks when 1 is heard
5	90% barks when 9 is heard, 10% barks when 8 is heard
6	20% barks when 2 is heard, 80% barks when 6 is heard
7	40% barks when 4 is heard, 30% barks when 0 is heard, 30% barks when 5 is heard
8	30% barks when 7 is heard, 20% barks when 9 is heard, 50% barks when 0 is heard
9	80% barks when 5 is heard, 10% barks when 1 is heard, 10% barks when 8 is heard
10	40% barks when 0 is heard, 20% barks when 5 is heard, 40% barks when 2 is heard

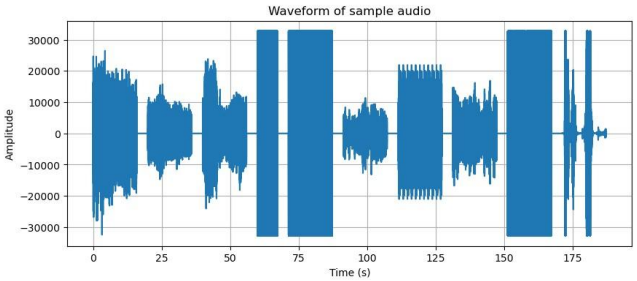

Figure 3: An example sound clip of a dog barking and different environmental sounds

Since each audio file is of a given length, feature extraction becomes challenging due to the vast amount of data involved. To tackle this issue, we segment each audio file into smaller pieces and mark the class as 'N' for the segments that the dogs did not react to as shown in Figure 4. This technique helps to reduce the size of each data fragment while increasing the amount of data used for extraction of waveform feature graphics. With this data augmentation method, the input data of 10 audio clips per dog is increased to approximately 500, allowing enough input data to support the training of the model.

Using the model's prediction results, each piece of data is stored as a predicted class in a local CSV file that forms the input data for the second model.

As noted, SVM, Decision Tree, Random Forest, Logistic Regression, and CNN were chosen as the primary models for the binary classification task [11]. SVM is a linear classifier often used for binary classification. It maps each instance representation to a

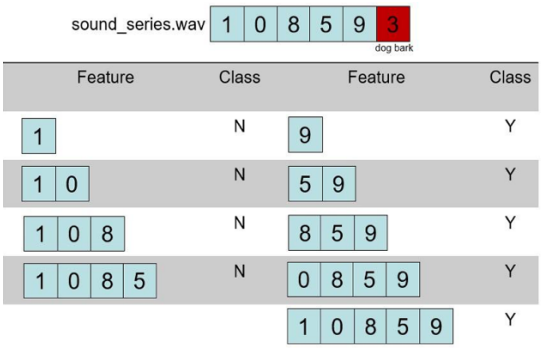

Figure 4: Data augmentation strategy

point in a two-dimensional space and decides the predicted class based on which side it falls. The input here is a series of sounds that a dog may hear. Since SVM is not based on probabilities, it only takes the multi-dimensional data as input and predicts a value that is binary, which makes it suitable for the task.

Decision trees build a multi-layer tree structure to determine the value of a given input. They are well suited to handle data with a given set of sound triggers without the need for excessive data pre-processing. Random forest models consist of multiple decision trees. They have an advantage over decision trees in that they can handle nonlinear features and adopt an approach based on voting for the highest result, i.e., whether a dog will bark is based on the model with the highest probability.

Compared to the above models, logistic regression provides a binary prediction task that predicts the probability that a dog barking event will occur. Such a probabilistic model is well suited for binary predictions as we want to be able to predict and thus prevent barking behaviour.

A neural network can handle a wide variety of input data types, including text, image, sound etc. A neural network generates a probability based on given input data where the barking probability is converted into a binary classification result to establish barking prediction.

We use the processed data in the above models separately and adjust the parameters accordingly to select the final model used for barking prediction on the 10 representative dogs presented in Table 2.

5 RESULTS AND ANALYSIS

As described, the final CNN model was based on a multi-input model consisting of two paths for the MFCC and reduced-dimension Mel Spectrogram elements. We use 10-fold cross-validation to evaluate the model to avoid degrading the generalization performance of the model, e.g., due to challenges of over-fitting. For each fold, the accuracy is between 60-80% as shown in Figure 5. We observed that the loss function decreases rapidly in the first ten epochs while the accuracy rises to over 80%. The model converges quickly and learns slowly after the 10 epochs. This means that both the loss function and learning rate have been chosen appropriately.

The Precision, Recall and F1 scores of the model over the entire 50 classes of audio files from the UrbanSound8K data sets shown

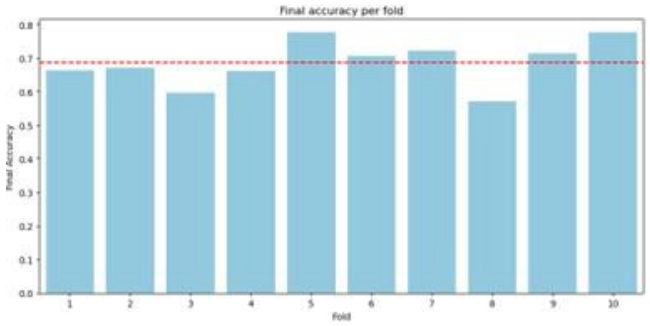

Figure 5: Accuracy over 10-fold cross-validation for 10 dogs barking behaviours

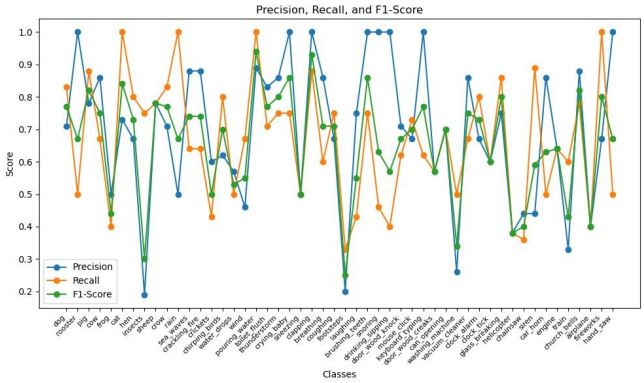

Figure 6: Precision, Recall and F1 Score over 50 Sound Classes from the UrbanSound8k Data Set

in Figure 6. This gives an overall representation of the prediction outcome of the audio features represented in Figure 1. We can observe that for specific noises such as helicopter and siren sounds, they have very similar Mel Spectrogram features, which is reflected in their relatively higher recall scores on the classification task.

Most importantly, it can be noted that the model the ability to classify dog barking with an overall accuracy of 77%. This is an essential component to predict the onset of sounds that might trigger dog barking. We note that the accuracy of natural soundscapes and water sounds is slightly lower compared to other categories. This might be a result of the acoustic features of these sounds being a bit more scattered and not as immediately obvious, i.e., there is no obvious feature discrimination between such sounds, hence there may well be classification errors. In addition, we found that for some classes the model had low accuracy but high recall, e.g., for insect sounds, indicating that the model may misrepresent other sounds as insect sounds. Overall, the environmental sound classifier reveals the ability to successfully classify the sounds of the environment with an accuracy of 66%. Considering that the task is to classify 50 different classes of sounds, the model provides a useful classification baseline that can be used for subsequent prediction tasks.

For each dog, we trained each model and tested the accuracy individually, and averaged the accuracy across all 10 dogs to evaluate

Table 4: Binary prediction task result

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.76	0.28	**0.31**	**0.28**
Logistic Regression	**0.85**	**0.66**	0.12	0.2
Neural Network	0.83	0.42	0.21	0.27
Random Forest	0.78	0.11	0.07	0.08
SVM	0.84	0.22	0.03	0.06

Table 5: Dog group barking prediction classification.

Model	Dog group	Accuracy
Logistic Regression	1-3	0.876
	4-6	0.874
	7-10	0.812

the overall performance of the model. As seen in Table 3, we found that logistic regression achieved the highest degree of accuracy (85%).

Considering complexity versus accuracy, even though the neural network had a relatively high accuracy (83%), neural networks are time-consuming models to train and incur more computational cost. Also, considering that the nature of the prediction task is to predict and hence prevent dog barking through treats, we prefer to predict every dog barking sound, i.e., a degree of false positives may well be acceptable [7] – and certainly would be acceptable to the dogs themselves! In this case, a lower precision score is acceptable compared to a lower recall rate. Compared with SVM and logistic regression models, logistic regression not only has a higher recall but also a higher accuracy and precision score. Therefore, from our experiments we can state that logistic regression is the most suitable model for the task of predicting dog barking.

Furthermore, as shown in Table 4, the accuracy decreases slightly as the dog's behaviour pattern becomes more complex. This is understandable since the complexity of the behavioural model implies a degree of randomness in the dog's behaviour. However, the model still has an accuracy of 81% for such complex dog behaviour triggers, i.e., for dog groups 7-10 from Table 2.

It is noted that for training and research purposes, the sounds used in the second model are from UrbanSound8k and these do not overlap with the ESC-50 sounds used for training. However, the classes of the sounds in ESC-50 and UrbanSound8k are not the same. For example, the gunshot class in UrbanSound8k is not present in ESC-50. This affects the prediction to some extent because there is no corresponding class to classify, e.g., gun sounds may be classified as fireworks. However, we believe that even the most comprehensive dataset cannot cover all sounds, and the same category may have completely different sounds associated with it. Capturing the acoustic characteristics of sounds is more important, since dogs may respond in the same manner to similar acoustic sounds. Thus, a dog may well bark in response to a gunshot, a firework or a car backfiring.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we explore whether it might be possible to avoid dog barking through positive reinforcement based on audio sounds that precede a dog barking. We construct a CNN-based multi-class classification model for detecting environmental sound events that are typically associated with home settings likely to trigger dog barking. Using the ESC-50 dataset, we propose Mel Spectrogram features and MFCC features for sound files and construct a CNN model with two input paths - this achieves an accuracy of 66%. We then use a collection of UrbanSound8k sounds as the basis for predicting dog barking behaviour. Utilising data augmentation and applying a range of machine learning approaches, we identify that logistic regression achieved a binary prediction model and achieved over 87% accuracy on simple audio triggers and over 81% for more complex dog behaviour triggers.

We have applied existing machine learning techniques to understand and predict dog barking behaviour which is a novel area of application. This work can be further refined and extended to deal with richer real-world scenarios and data, as well as integration with IoT devices for dispensing of treats.

Whilst we have successfully demonstrated the ability to accurately predict barking behaviour, the work is not without some limitations. Firstly, the input audio model data relies on two datasets: ESC-50 and UrbanSound8k. Even though these two datasets have a large amount of data that is used to underpin a body of research, they cannot cover all factors that can trigger dog barking in real life. The triggering sound varies from dog to dog due to differences in living conditions. The dog barking prediction model strongly relies on the specific environmental sounds connected to each household. Adapting the models and/or using transfer learning to facilitate the real-world deployment of the models would be an obvious extension to the work.

A further limitation is that there is no available real-world dataset that can be used for the dog bark prediction model, hence a synthetic audio sample was used to simulate the sounds that "might" be collected in real life based on different audio patterns and representative dog behaviours. The actual real-world situation is clearly more complex with ambient noises that occur continuously through the day or external noises not included in the training data set. Thus, even if we use data augmentation to increase the amount of available data, it is difficult to represent the richness and diversity of real-world data.

Another challenge is that it is very difficult to build a large and complex CNN model considering the time cost and the training complexity. The models above are ultimately intended to be deployed to IoT devices to activate a pet treat dispenser. Vish [5],

OpenL3 [8], and Yanet [12] use lightweight neural network models targeted to sound based scenarios that have been shown to perform well and can be considered in future extensions to the work.

It is also worth mentioning that dog barking is triggered by a variety of reasons, and sound may be only one factor. The dog's personality, upbringing, and environment may all influence the dog's sensitivity to sound. For example, a large dog or an older dog may be gentler and less responsive to louder sounds, whilst a smaller dog or a dog that has been neglected or abused may be more sensitive to noise. Considering the variability among dogs, the generalization ability of a single model to cover all scenarios is challenging. In addition, a single sensor cannot analyse a dog's real-time emotional state, physical health, and immediate environment, which may all have an impact on the dog's tolerance to specific sound combinations.

There are also ethical considerations that need to be fully considered, both when gathering the data and in the training phase, e.g., to ensure pet owner's privacy.

There are several possible future research directions to this work. The first thing that can be improved is the collection of data. A large amount of data can be collected on different breeds of dogs, their size, and the different home environments. With more available data, different models can be used to test on the dataset to understand dog behaviour patterns and the response to different sounds.

The introduction of real-time acoustic event detection and unsupervised learning can also be considered. In this way, when a dog hears a specific sound, the system will detect the acoustic event in real-time and give timely feedback to train the dog's behaviour in real-time. At the same time, unsupervised learning can continuously learn the sounds around the dog and adjust its own prediction and feedback according to the real-time changes and the dog's behavioural response.

More sensors and analytics can also be employed. For example, the visual information provided by other sensors such as cameras can be added to analyse the environmental situation. For example, strangers who make no sound and the passing of other dogs may also act as a trigger to the dog to start barking.

The emotional analysis of barking sounds and body movements can also be added. If the emotional state of the dog can be determined, it may be possible to calm the dog and train positive behaviour more effectively. The classification of the emotion and mood of pets was explored in [27]. Other dimensions of animal behaviour could also be explored and included, e.g., the difference between an angry/aggressive barking behaviour and a scared behaviour. Examples such as [28] have considered such other broader dimensions in other species, e.g., whether a cat might be in pain or not.

Furthermore, information about the tendencies of specific breeds could be incorporated to create more accurate models for predicting when a dog might be about to bark.

Finally, we can develop models that are targeted to specific households. Thus, the models developed in this paper are based on a set of noises reflecting what might be heard in a given home setting using a representative collection of dogs with different audio triggers. Clearly, this is the proof of concept and in the real world, there will be many other localised noises that act as triggers for each species and instance of dog.

ACKNOWLEDGMENTS

The authors would like to thank veterinary scientist Dr Dennis Wormald for the initial ideas behind this work. This research was undertaken using the LIEF HPC-GPGPU Facility (SPARTAN) hosted at the University of Melbourne. This facility was established with the assistance of Australian Government LIEF Grant LE170100200.

REFERENCES

[1] A. Bansal and N.K. Garg. Environmental sound classification: A descriptive review of the literature. *Intelligent Systems with Applications*, page 200115, 2022.
[2] A.L.A. Barber, A. Wilkinson, F. Montealegre-Z, V.F. Ratcliffe, K. Guo, D.S. Mills. A comparison of hearing and auditory functioning between dogs and humans. *Comparative Cognition & Behavior Reviews*, 15:45–94, 2020.
[3] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.H. Tauchert, and K.H. Frommolt. Detecting bird sounds in a complex acoustic environment and application to bio-acoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010.
[4] S. Chachada and C.C.J. Kuo. Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, 3:e14, 2014.
[5] N. Di, M.Z. Sharif, Z. Hu, R. Xue, and B. Yu. Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification. *PeerJ*, 11:e14696, 2023.
[6] S. Furui. History and development of speech recognition. *Speech Technology: Theory and Applications*, pages 1–18, 2010.
[7] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, pages 345–359. Springer, 2005.
[8] S. Grollmisch, D. Johnson, J. Abeßer, and H. Lukashevich. Iaeo3-combining openl3 embeddings and interpolation autoencoder for anomalous sound detection. *Tech. Rep., DCASE2020 Challenge*, 2020.
[9] E.F. Hiby, N.J. Rooney, and J.W.S. Bradshaw. Dog training methods: their use, effectiveness and interaction with behaviour and welfare. *Animal welfare*, 13(1):63–69, 2004.
[10] M.M. Kabir, M.F. Mridha, J. Shin, I. Jahan, and A.Q. Ohi. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*, 9:79236– 79263, 2021.
[11] K. Kumari and S.K. Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7), 2017.
[12] C. Malmberg. Real-time audio classification on an edge device: Using Yamnet and Tensorflow lite, 2021.
[13] Molnar, F. Kaplan, P. Roy, F. Pachet, P. Pongracz, A. Doka, and A. Miklosi. Classification of dog barks: a machine learning approach. *Animal Cognition*, 11:389–400, 2008.
[14] G. Muhammad and M. Melhem. Pathological voice detection and binary classification using mpeg-7 audio features. *Biomedical Signal Processing and Control*, 11:1–9, 2014.
[15] K.J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.
[16] P. Pongracz, C. Molnar, and A. Miklosi. Barking in family dogs: an ethological approach. *The Veterinary Journal*, 183(2):141–147, 2010.
[17] Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z.A. Kaminsky. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, 3(1):78, 2020.
[18] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.
[19] R.N. Singarimbun, E.B. Nababan and O.S. Sitompul. Adaptive moment estimation to minimize square error in backpropagation algorithm. In *2019 International Conference of Computer Science and Information Technology (ICoSNIKOM)*, pages 1–7. IEEE, 2019.
[20] D.B. Springer, L. Tarassenko, and G.D. Clifford. Logistic regression-hsmm-based heart sound segmentation. *IEEE transactions on biomedical engineering*, 63(4):822–832, 2015.
[21] A. Statnikov, L. Wang, and C.F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):1–10, 2008.
[22] Y. Su, K. Zhang, J. Wang, and K. Madani. Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors*, 19(7):1733, 2019.

[23] S.S. Talathi and A. Vartak. Improving performance of recurrent neural network with ReLU nonlinearity. *arXiv preprint arXiv:1511.03771*, 2015.
[24] K. Tiira, S. Sulkama, and H. Lohi. Prevalence, comorbidity, and behavioral variation in canine anxiety. *Journal of Veterinary Behavior*, 16:36–44, 2016.
[25] J.C. Wang, H.P. Lee, J.F. Wang, and C.B. Lin. Robust environmental sound recognition for home automation. IEEE transactions on automation science and engineering, 5(1):25–31, 2008.
[26] X. Wang, X. Zhao, Y. He, and K. Wang. Cough sound analysis to assess air quality in commercial weaner barns. *Computers and Electronics in Agriculture*, 160:8–13, 2019.
[27] R.O. Sinnott, U. Aickelin, Y. Jia, E.R.J. Sinnott, P.Y. Sun, R. Susanto, Run or Pat: Using Deep Learning to Classify the Species Type and Emotion of Pets, *IEEE Conference on Computer Science and Data Engineering, Gold Coast, Australia*, December 2021.
[28] Y. Yang, R.O. Sinnott, Automated Recognition and Classification of Cat Pain through Deep Learning, *IEEE DataCom conference, Fiji*, December 2022.