### Abstract

Excessive and intense barking has been a long-time problem for dog owners and households due to the harshness of the sound, the loud acoustic nature of the volume, and the fear of barking by a significant proportion of the population. The causes of dog barking are varied however barking caused by environmental noise has been particularly noticed. With the maturity of sound recognition technology, it has become possible to represent sound waveforms from Mel-frequency cepstral coefficients (MFCC) and Mel Spectrogram features to model and predict the time of appearance of barking dogs by a Convolutional Neural Network structured model. This paper will perform a classification model on possible dog bark triggers from public environmental sound databases and use the model to analyse and predict the barking behaviour from audio clips simulating a real-life dog barking scenario.

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

### I. Introduction

**- Background information on the project**

In recent years, with the rapid development of deep learning, sound recognition has also been widely used in vary or fields, including automatic speech recognition (ASR)\cite{kabir2021survey} and environmental sound classification (ESC) \cite{bansal2022environmental}, and has been widely applied in smart home, medical, security monitoring fields. The emergence of audio recognition based on the development of machine learning and neural networks has made it possible to handle complex audio classification tasks. In addition to the areas mentioned, there has been quite a wide range of applications for animal voice recognition. The convolutional neural network (CNN) models outstanding ability to distinguish differences in audio of has made it possible perform breed identification and environmental quality issue based on animal sound analysis\cite{wang2019cough}.

Large numbers of automation of decisioning and classification of sound-based processing tasks have been raised and discussed. In this paper, a method to predict the cause of dog barking based on environmental sound recognition will be demonstrated. Due to the large variety and amount of noise in the city, these noises would cause anxiety in dogs and leads to excessive barking. Each individual dog is sensitive to different noises and lead to different reactions, as a result each dog will have their different noises that trigger barking depending on personality and environment. It is therefore natural for dog owners to seek to train their pets to minimize the behaviours that cause discomfort.

This prediction task of dog barking the following difficulties: (1) the nature of large amount and variety of urban noises around households therefore for general classification of noised a dataset contains large range of different noised is required; (2) difficulty to collect a large amount of data of each individual as the barking trigger sound is different, and there is almost no audio datasets about dog trigger barking is available, hence audio data augmentation technique is important; (3) based on breed, characteristic, growing environment etc. each dog would has different behavioural patterns, a general strategy to cover all to predict barking behaviour for a specific dog; (4) major limitation of the model about barking behaviour varies from each independent dog, and sound recognition could hardly cover all sources of barking triggers.

This paper discusses a machine learning-based approach to automate the prediction and barking behaviour of domesticated dogs to achieve behaviours that can intervene and prevent barking triggered by environmental noise and could further positively motivate the correct behaviour of dogs. Section 2 will go through a literature review of audio recognition, environmental sound classification systems, dog barking veterinary behaviour, and existing datasets; section 3 introduces the methodology used in the research, comparing different data processing, feature extraction, data augmentation techniques, CNN models and pre-trained model and hyperparameter tuning methods, managing to achieve a better performance in accuracy in efficiency; section 4 would evaluate the experiment result and demonstrate the model performance; section 5 discusses the limitation and potential future directions; and section 6 concludes the work and research. By conducting this research, an automated system for dispensing treats or snacks to the dog before it barks and using the model as a tool to encourage positive behaviour is aimed to be achieved.

### II. Literature Review

**-2.1 Techniques for sound classification**

The earliest research into voice recognition was scholars began to explore automatic speech recognition (ASR) in 1950s. Due to limitations in computing power as well as lack in sound analysis technology, the main methodology was pattern matching based on acoustic principles \cite{furui2010history}. The first big breakthrough of sound recognition was Hidden Markov models (HMMs) methods in the 1980s. The nature of excellent statistical power of HMMs models make it suitable and widely used in sound classification tasks. Later in the 1990s, the exponential increase in computer computing power and the widespread availability of audio files led to an explosion in automatic speech recognition including large vocabulary continuous speech recognition (LVCSR) systems. Now due to the advancements in deep learning in recent years, neural networks have demonstrated exceptional suitability for recognition tasks in effectiveness and accuracy, Support Vector Machines (SVMs) and CNN have become the predominant approach for voice recognition\cite{su2019environment}.

**- 2.2 Overview of environmental sound classification**

Research on environmental sound classification has been largely focused around the last 20 years when machine learning and neural networks have been widely used. Speech recognition and music classification are the initial focus of attention on environmental sound recognition. Later the improvement of feature extraction in time domain features, frequency domain features, and time-frequency features like Mel Frequency Cepstrum Coefficient (MFCC) brought in the development of environmental sound classification\cite{chachada2014environmental}. Recently, environmental sound classification is widely used in home automation\cite{ wang2008robust} and in the field of sound-based species classification such as bird call classification\cite{ bardeli2010detecting}.

**- 2.3 Binary classification of sounds**

There is also extensive published work on the binary classification of sounds, with popular models including Support Vector Machine (SVM)\cite{muhammad2014pathological} and the CNN as mentioned above. In addition, numerous studies on biological characteristics classification have cited the effectiveness of random forest trees\cite{statnikov2008comprehensive} and logistic regression\cite{springer2015logistic} in similar classification tasks.

**- 2.4 Previous studies on dog behaviour and barking**

In existing research of dog behaviour, dog barking is not only a reaction to express anger, but also for many reasons. Studies have shown that dogs will bark when faced with fear, noise and separation anxiety, and female, neutered and more fearful dogs significantly responding to noise more strongly\cite{tiira2016prevalence}.

In addition, dog barking has been shown to have a role in communication. When dogs bark in communication and other behaviours, dog barking will be different from other dogs, and it will be affected by human preferences and social environments\cite{pongracz2010barking}.

**- 2.5 Existing datasets for environmental sound classification**

With the popularization of digital recording technology and more and more sound file could be retrieved from the internet, researchers began to construct large-scale environmental sound datasets. These datasets include different types of environmental sounds such as city streets, forests, traffic noise, etc. In recent researches the two databases widely used for urban environmental sound classification are UrbanSound8K\cite{Salamon:UrbanSound:ACMMM:14} and ESC-50\cite{ piczak2015dataset}. Two databases consist of a collection of urban sounds, which includes 8,732 common sounds heard in a city that also contains dog bark. Waveform signals will be extracted from all sound files for feature extraction. Researchers could classify and analyse environmental sound events based on these datasets and apply the models to detect specific sounds such as fire alarms, car horns, etc.

### III. Methodology

Environment sound classification model

The high-level approach includes a multi-class classifier and developed using a CNN-based neural network model. The emphasis of this model will be on using appropriate convolution kernels to efficiently extract waveform signal features. To achieve the best training results for this classification task, various convolutional neural network structures, data inputs, number of pooling layers, and learning rate parameters will be explored.

**- Data collection and pre-processing**

As a model for detecting environmental sound events, it is necessary to be able to classify a wide range of different classes of environmental noise. The dataset used for the first ESC classification task is ESC-50, which has 50 types of common urban sounds from 5 higher-level categories: Animals, 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 the home environment thus is suitable to be the data source to train the classification model of dog bark triggers.

The dataset contains an csv file and 2000 wav format audio clips to process with. We use LibROSA, a python library to read the wav files into a one-dimensional array format of audio samples time series representation of the audio with the sample rate.

**- Data augmentation and regularization techniques**

**Balance the dataset -** In the ESC-50 dataset, we have a total of 2000 audio files, with each category contains 40 samples. Each of these audio files has a duration of 5 seconds. This ensures an evenly distributed representation across all categories hence no further balance process is required.

Data normalisation - For each audio, we normalise the waveform by standard deviation and centre it to the zero mean. This ensures that over all audio files, the volume level is consistent and clipping and any distortion can be avoided. This process also minimises the difference between audios in the same category caused by different radio conditions and outstands the differences in acoustic characteristics between different classes.

Input length - Table lists out the different audio training input lengths experimented. On a base CNN model, we tried separately using (1) original 5s long audio (2) normalized 5s long audio (3) normalized random 2s long audio (4) normalized random 2s long audio, selected non-silence interval (5) normalized random 3s long audio, (6) normalized random 3s long audio, selected non-silence interval, extracts the basic MFCC features as input. By comparing the accuracy, the normalized 5s long audio is used as the base audio input.

Non-silent intervals - We used the function librosa.effects.split() to split all non-silent segments, but it also captures small nonsensical segments that is meaningless to the training. We believe that for some human sounds with intermittency acoustic features, such as coughing and snoring, shorter duration reduces the model's ability to discriminate audio. Although this reduces the amount of input data and training time, the outcome suggests that the silence gaps in the audio contributes positively to distinguish them from other sounds.

**- Feature extraction methods**

As mentioned in section 2, both MFCC and Mel Spectrograms are the dominant approach of extracting features from waveform signals.

Mel Spectrograms provided a visual representation of the sound spectrum. The original audio signal is transformed into a spectrogram by Mel Spectrogram, which shifts audio information from the time domain to the frequency domain, where the vertical axis represents frequency, the horizontal axis represents time, and the colour or grayscale represents energy or amplitude.

MFCC is a standard technique in audio recognition and is set of eigencoefficients extracted from the Mel spectrogram. It captures characteristics such as pitch, timbre, and other important properties related to human perception of sound. The MFCC extraction process includes pre-processing, fast Fourier transform, Mel scale, logarithmic operation, discrete cosine transforms, dynamic feature extraction and other steps, smoothing of the spectrum by reinforcing the wave at low frequencies as in human aural. Although it is commonly believed that dogs hear farther and in a larger frequency range than humans\cite{barber2020comparison}, the main target is to distinguish environmental sounds and MFCC’s ability of compress wave feature dimension of makes it well suited in classification task.

The graph below gives a visualised view of the Mel Spectrograms and MFCC. The Mel Spectrogram represents the time and frequency characteristics of the waveform well and is well differentiated between each different class. The MFCC coefficients on the right also gives a fair representation of acoustic characteristics of each class even though the dimensionality is reduced. Intuitively MFCC conveys the waveform features well on the low frequency area.

We trained the basic CNN model by using both features. As a preliminary result, using MFCC feature only gives an accuracy around 63% and using Mel Spectrogram gives 57%. Comparing with the accuracy MFCC and Mel Spectrograms used as the feature separately, and considering in balancing the training efficiency and accuracy, in the final model both features were used as the feature to train the model. While MFCC gives a good representation, we still want to keep the higher frequency areas as a feature as dog are sensitive with them. Both features are extracted from the pre-processed video clip and saved as npy (NumPy array format) locally for further use to save time for further training process.

**Convolutional layers details**

The benchmark CNN used was a multi-input model that has two branches built in with. The MFCC features are processed on one path, and on the other path Mel Spectrogram features are processed. The input data were handled individually by each branch using convolutional, pooling, and fully connected layers. The pooling layers reduce computational complexity and feature dimensionality. The outputs of the two branches were then combined and fed through a dense layer for final processing and prediction.

Mel-Spectrogram Path – The raw Mel Spectrogram feature has a dimension of (128, 216). To save the training time, a MinMaxScaler to scale Mel spectrogram data to the [0, 1] range is used to reduce the dimension of the features. This also speeds up the time of the model to converge by minimise the interval of the feature and prevent large values from impacting the model. Then ReLU activation function is used to process the Mel spectrogram input through two 2D convolutional layers with 16 and 32 filters, and max pooling is used to lower the spatial dimensionality after each convolution layer. The output is then flattened to connect to a Convolutional Autoencoder for reducing dimension. The Convolutional Autoencoder employs Conv2D, MaxPooling2D, and UpSampling2D layers. A fully linked layer with an encoding dimension of 50 and a ReLU activation function in the encoder processed the output from the above steps. Then the encoded Mel-spectrogram data uses a Dense layer to process the encoded Mel spectrogram data to reduce the weight in the combined CNN model.

MFCC Path – The MFCC path is used to process the MFCC features, which are the main features used in our classifier. It is 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 .

Concatenation and Dense Layers - A Concatenate layer is used to combine the output of the MFCC path with the Mel spectrogram path and follows with a dense and employs a SoftMax activation function to execute the classification of environmental sounds.

**Hyperparameter tuning**

**Convolutional Layers (Conv2D)** - We employ a convolutional kernel of size (3, 3), which is the most typical size for convolutional neural networks in the convolutional layer. This size enables efficient and detail-preserving waveform feature capture. The number of filters affects the complexity of the CNN learning to the features. For the first convolutional layer32 filters used to interpret the basic time & frequency features in MFCC, and then two layers of 128 filters are used to capture the more detailed characteristics of the MFCC feature. The activation function here used is Rectified Linear Unit (ReLU), it is computationally efficient to save the training time and also works well on avoiding the vanishing gradient problem\cite{talathi2015improving}.

**Pooling Layers (MaxPooling2D)** – 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 are set to (2,2) to reduce the complexity to approximately half size as the output improve the efficiency of training.

**Dropout Layers -** The dropout layer is used to avoid over-fitting the model, in other words to improve generalization. In our model dropout rate is set to 0.3 that is 30% of input unit has been set to zero, effectively reduces interdependencies between the input units.

**Dense Layers –** The number of neurones in the fully connected layers is set to 128 and 512 which is appropriate for MFCC features, and output layer has 50 neurons with 'SoftMax' activation as the classification task has 50 class to classify as the SoftMax function transforms the output of a into a probability distribution.

**Model Compilation** - Adaptive Moment Estimation algorithm (adam) optimization is a popular optimization algorithm used in deep learning model. 'Adam' has adaptive learning rate and an efficient parameter update algorithm; it effectively updates the weight between each layer of the neural network and makes it widely applicable in neural network models. Lost function used is 'categorical\_crossentropy’, which is widely used in multiclass classification problem. Cross-entropy loss of the predicted class probabilities and the true category labels is calculated in each batch. Categorical cross entropy generates smooth and continuous gradients in the backpropagation process, assisting in the reduction of loss when employing the stochastic gradient descent (SGD) approach.

**Model Training –**Considering the training efficiency the batch size is set to 8 for quicker weight updates and potentially faster convergence. Number of training epoch has been set to 40, which is set as a result of multiple training sessions to ensure enough training cycles while preventing overfitting of the model.

Dog barking behaviour prediction model

The second classification task is binary classification of audio, 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. Gathering sufficient data for each dog is time-consuming and impractical. Overcoming these challenges requires approaches such as data augmentation. Despite the limitations, advancements in machine learning techniques offer opportunities to develop effective dog behaviour training models in real-life scenarios. Addressing these challenges can contribute to understanding and predicting dog behaviour more effectively. Since each dog may behave differently, the training data will contain combined sounds based on UrbanSound8k. Separate noises and barks are combined into longer audio files using audio clips. Multiple models will be built and compared for this classification task to select the best performing model.

In the following section, SVM, Random Forest, Logistic Regression, and CNN are experimented with and contrasted. Also, data augmentation, regularization, and hyperparameter tuning techniques would be balanced likely improving the performance of the model and achieving a better result.

* **Data collection**

The data used to predict dog barking behaviour was UrbanSound8K dataset, an aggregation of 8732 common urban sounds over 10 classes, includes dog barks was used. We believe that this dataset is suitable for the triggering dog barking task as the 10 categories covered in the dataset: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music are all sounds that are commonly heard by domesticated dogs and cover a wide range of acoustic properties.

As analysed earlier, it is clearly unrealistic to use the same trained model to match the barking behaviour of the dog, since each dog reacts differently to each sound. Therefore, the approach would be finding a suitable model and training the model individually for each dog. To address the difficulty mentioned in the previous section about collecting sound samples from each dog, we designed a method that can be practically applied in life: simulating the real-life collection of sound samples from barking dogs. To simulate the data obtained in the real scenario, we first designed the behaviour pattern of 10 dogs as listed in the table:

As shown in the figure, each dog has their own individual behaviour patterns. We used the audio samples from the UrbanSound8K dataset, following each dog's behaviour pattern to splice the sound 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.

**Data augmentation**

Since each audio file is lengthy, feature extraction becomes challenging due to the vast amount of data involved. To tackle this issue, our initial approach is to segment each audio file into smaller pieces and mark the class as ’N’ for the segments that the dogs didn’t react with. This technique will help reduce the size of each data piece while increasing the amount of data to facilitate the extraction of waveform feature graphics.

**Prediction methodology**

As the audio clips are generated for each dog to simulate the sounds that would be heard and place them in the model for environmental sound detection. Using the model's prediction results, each piece of data is stored as a predicted class in a local CSV file as the base data for the second model.

**Model evaluation and comparison**

SVM, Decision Tree, Random Forest, Logistic Regression, and CNN have been chosen as a suitable model for the binary classification task. We ran the data to through a baseline model to test the accuracy of different models for this bark prediction and evaluate and pick the most appropriate model.

SVM - SVM is a linear classifier for binary classification. SVM maps each instance representation to a point in a two-dimensional space and decides the predicted class based on which side it falls on. While our input is a series of sounds that a dog may hear, as SVM is not based on probabilities, it would only take the multi-dimensional input and predict a value that is binary, which makes it suitable for the task.

Decision Tree - Decision trees build a multi-layer tree structure to determine the value of a given input. The reason why decision trees are suitable for this model is that they are well suited to handle data with a given set of sound triggers without much data pre-processing.

Random Forest - The random forest model is a model consists of multiple decision trees. This model has an advantage over decision trees that it can handle nonlinear features and vote for the highest result, which means that whether it will bark will depend on the model with the highest probability.

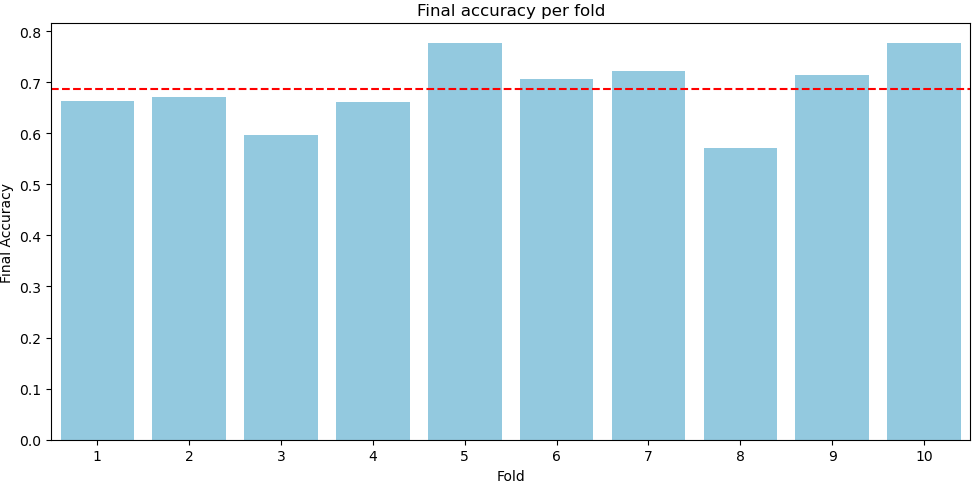
Logistic Regression - Compared to the above models, the advantage of logistic regression model is logistic regression is a statistical learning-based model, in this binary prediction task it predicts the probability rather than deciding an outcome that a dog barking event will occur. This probabilistic model is well suited for binary predictions as we want to be able to predict and prevent the barking behaviour.

Neural network - is broad used as it can handle a wide variety of input data types, including text, image, sound etc. generates a probability based on the input data and the barking probability is converted into a binary classification result like logistic regression, which is ideal for our barking prediction.

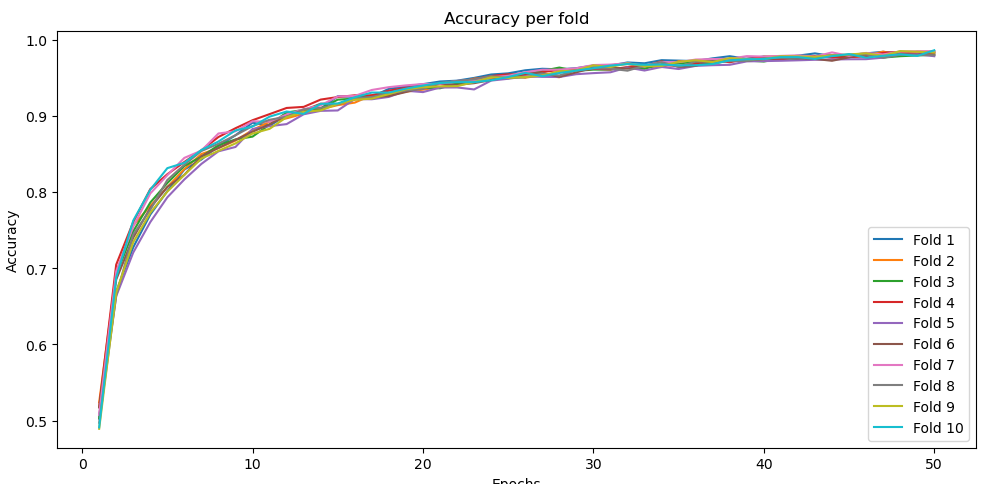
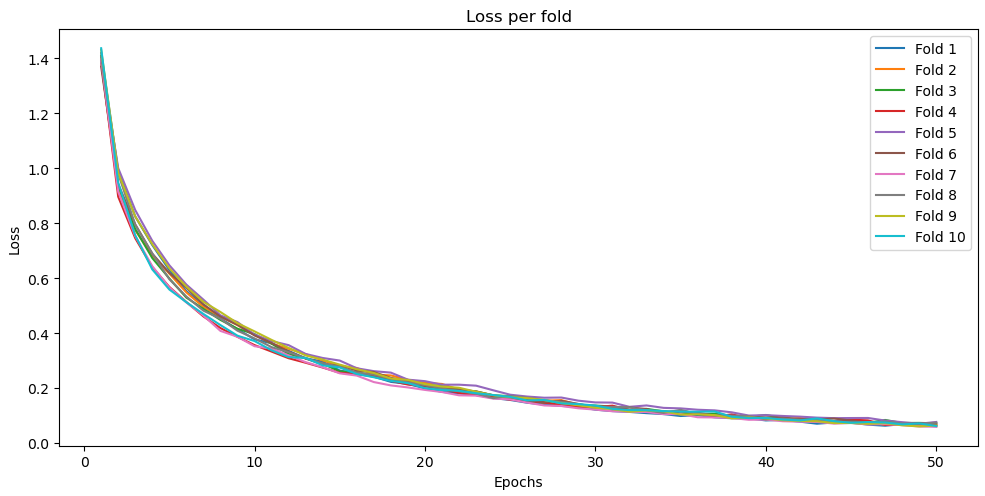
### IV. Results and Discussion

**- Environmental classification task result**

The final CNN model used is a multi-input model consisting of two paths with mfcc and reduced-dimension mel-spectrogram.



For each fold, the accuracy is between 60-80%. The model has evaluated more accurately using the 10- cross-validation to avoid degrading the generalization performance of the model due to over-fitting.

We used cross-validation to evaluate the model over 10 folds. We found that the loss function decreases rapidly in the first ten epochs while the accuracy rises to over 80. The model converges quickly and learning slowly after the 10 epochs, which means that our both loss function and learning rate has been chosen appropriately.

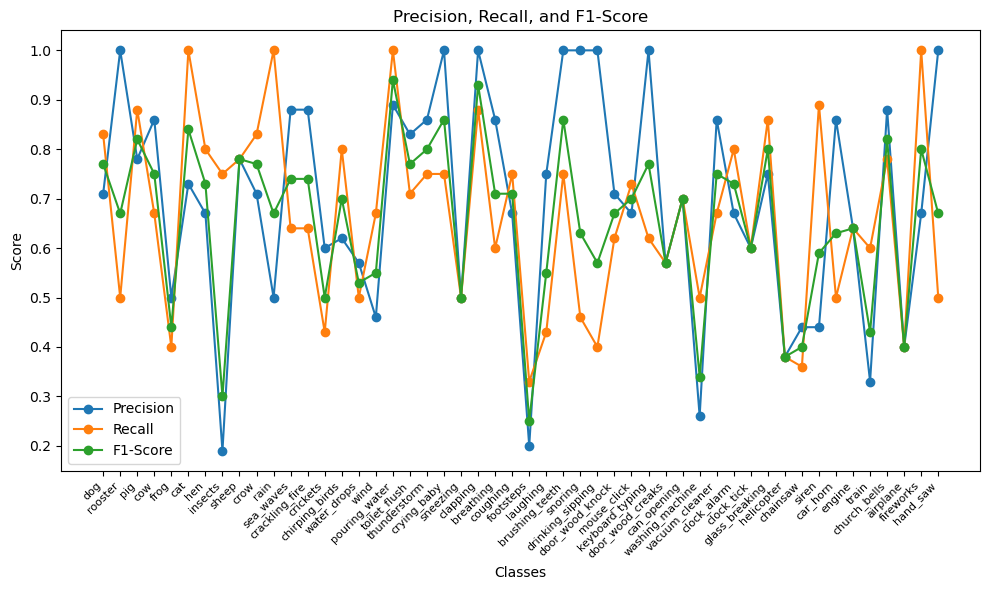
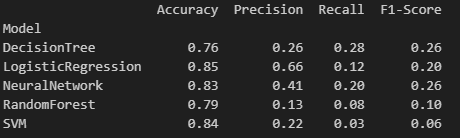


Figure gives the Precision, Recall and F1score of the model over the 50 classes. The first thing to note is the model's ability to classify dog bark. From the F1score of 0.77. it is to say that the model can recognize dog barking well, which is necessary for the subsequent sound collection and prediction. We note that the accuracy of Natural soundscapes & water sounds is slightly lower compared to the other major categories. This might be a result in the acoustic features of these sounds are a bit more scattered and do not have obvious This may be since the acoustic features of these sounds are a bit more scattered and there is no obvious feature discrimination leading to classification errors. In addition, we found that for some classes the model has low accuracy but high recall for bug sounds, indicating that the model often misrepresents other sounds as bug sounds.

Overall, the environmental sound classifier reveals the ability of successfully classifying the sounds of the environment of an accuracy 66%. Considering that the classification task is to classify in 50 different classes of sounds, the model could provide useful classification and be helpful for the subsequent prediction tasks.

**- Binary prediction task section result**



For each dog, we trained each model and tested accuracy individually, and averaged the accuracy across 10 dogs to evaluate the performance of the model. By comparison on table, we found that logistic regression, support vector machines and neural networks are more accurate in terms of accuracy. Considering complexity versus accuracy, even though the neural network has a relatively high accuracy, neural networks is a time-consuming model which implies more computational cost. Also, considering the nature of the prediction task is to prevent dog barking by predicting dog barking, we prefer to predict every dog barking sound while false positives are acceptable. Therefore, it is necessary to analyse Precision and Recall. 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 higher recall but also higher accuracy and precision score. Therefore, we believe that logistic regression is the most suitable model for this task of predicting dog barking.

**- Evaluation of model performance over dog behaviour pattern**



From the table, accuracy decreases slightly as the dog's behaviour pattern becomes more complex. We believe that this is acceptable because the complexity of the behavioural model implies randomness to the dog's behaviour, but the model still has an accuracy of 81\% for complex dog behaviour (Dog 7-10).

**V. Findings and Discussion**

**- Limitations**

Although this model can successfully predict barking behaviour, there exists limitations and shortcomings in applying them into a real-life scenario.

First, our model data rely on two datasets, ESC-50 and UrbanSound8k. Even though these two datasets have a large amount of data as well as existing research, they still cannot cover all the factors that can trigger dog barking in real life The triggering sound varies from dog to dog due to differences in living conditions. Furthermore, the second dog prediction model strongly relies on the preceding environmental sound model's detection results. The prior model has a accuracy of 66\%, which will affect the accuracy of the subsequent dog barking prediction model. In addition, because there is no available dataset, for the dog bark prediction model we are using synthetic audio to simulate the sounds that might be collected in real life based on established behaviour patterns. The actual dog situation is more complex, and the specific data collection may be influenced by various factors such as radio equipment. 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 point is that it is very difficult to build a large and complex CNN model considering the time cost and the training complexity. Both of our models above consider the training time and computational power and reduce the complexity, including the dimensionality of the features and the complexity of the model. Even if we find a relatively balanced point, it is unavoidable to reduce the training accuracy due to the reduced complexity. VGGish, OpenL3, and YAMNet are all neural network models used to train sound and have been shown to perform very well.

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 may be gentler to louder sounds, or a dog that has been strayed or abused may be more sensitive to noise. Considering the variability among dogs, the generalization ability of one model may not be sufficient to cover all barking behaviours. In addition, only one sensor, sound, cannot analyse the dog's real-time emotional state, physical health, and immediate environment, which may all have an impact on the tolerance to a specific sound.

**- Recommendations for future research**

For future work, one direction is to involve real-time Acoustic Event Detection and unsupervised learning, that it could be able to train or manage a dog's behaviour in real time, changing based on the sounds around the dog.

Another is to combine veterinary behaviour studies to improve our model. If we know more about the tendencies of specific breeds, or health status of individual dogs, we could make a completer and more accurate model for predicting when a dog is going to bark.

**VI. Conclusion**

**- Summary of the study**

Overall, this research signifies a direction in the use of machine learning techniques to understand and predict dog barking behaviour, which could be further refined and applied in real world scenarios such as in training aids or behaviour management tools for dogs.