Acoustic Classification of Cat Breed Based on Time and Frequency Domain Features

William Raccagni
Università degli Studi di Milano
Milan, Italy
william.raccagni@studenti.unimi.it

Stavros Ntalampiras
Università degli Studi di Milano
Milan, Italy
stavros.ntalampiras@unimi.it

***Abstract*—The emerging field of Bioacoustics has been presenting significant research activity lately, and thanks to the use of machine learning methods, several tools and methodologies have been established for identifying certain patterns and meanings in animal vocalizations. Animal sounds can vary over time in intensity and patterns produced between different breeds of the same species, both for physiological reasons and for different emotional states and needs. Pets, such as dogs and cats, are no exception, thus allowing a vocal distinction between breeds. This article studies classification of the cat breed, in particular on the Maine Coon and European Shorthair breed, based on the public audio dataset "CatMeows". To this end, we employed features coming from time and frequency domain capturing relevant information as regard to the present audio structure. Subsequently, audio pattern recognition was carried out by means of k-means clustering, k-NN, and multilayer perceptron learning models. After extensive experiments, we obtained very promising results , with an average accuracy that runs around 98%. In particular, time-domain features presented a strong contribution, as demonstrated by the results using k-means.**

I. Introduction

During the last decade, the exciting developments in the machine learning field have paved the way to novel applications in the constantly growing field of bioacoustics, where audio pattern recognition plays a relevant role. The interdisciplinary scientific branch of bioacoustics analyzes and studies the production, dispersion and reception of sounds in animals [1], [2]. Moreover, the field of bioacoustics studies the sounds produced by the fauna of different ecosystems, in order to trace the habits of the animal species that compose it [3], [4]. Different animals produce sounds with different spectral patterns and intensities over time, based on the animal's anatomy and cognitive abilities [5]. The differences are not limited only to the species, but also to the different breeds that compose it. Constraining the problem in studying the differences between different breeds of cats [6], allows us to better investigate the differences between the vocalizations, thus identifying the best strategies for classifying the breeds of the same species. Therefore, identifying a correct strategy for the identification of different breeds would allow us to estimate and monitor the animal distribution in different types of ecosystems, in a non-invasive audio-based way.

The related literature includes automatic classification approaches of different animal vocalizations primarily aiming at physiological aspects (such as breed and sex), emotional and/or contextual aspects. The majority of these works mostly focuses on datasets encompassing cat and dog vocalizations. This is due to a) such pets are characterized by high availability and management, b) convenience in recording samples of such animals with good quality, and c) such animal-human relationships exist since thousands of years and has reached the level where research can be conducted towards interpreting the meanings behind each animal vocalization.

The literature encompasses a significant amount of research on the study and classification of dog barks. In [7], the authors studied the classification of sex, age, context and individual from vocalizations of Mudi dogs. The dataset developed by the authors is composed of 800 registrations of 8 dogs registered in different contexts. For each recording, 29 different features were extracted, 1 for each method/statistic. The classifiers used for the comparison are Naive Bayes, Classification Tree, k-Nearest Neighbors and Logistic Regression. Interestingly, the authors of [8] investigate the automatic classification of 5 dog emotional states. The "EmoDog" Dataset is composed of 226 bark sequences recorded by 12 different Mudi dogs. The feature sets studied are variants of the EGEMAPS and COMPARE feature sets [9]. Support Vector Machine (SVM) classification was applied to each set to classify the 5 emotional states. The work presented in [10] examined dog bark classification into 6 different contexts, i.e. *play*, *fight*, *alone*, *stranger*, *walk* and *ball*. The dataset used is a collection of 6646 barks of 14 Mudi dogs of different ages and sex, recorded in the 6 contexts mentioned above. A wide variety of time and frequency domain features was considered and filtered by selecting only a subset using a Naive Bayes based algorithm. The most significant features were employed to train a Bayesian classifier. In [11], automatic individual and breed classification of different dogs (Chihuahua, French Poodle, Schnauzer and Others) was studied. The dataset used and developed by the authors is made up of 6103 barks from 36 different individuals. Three different sets of features were used (IS-09, IS-10, IS-11 [9]), plus a fourth obtained through various feature selection methods. Six different classifiers were compared with respect to classification of dog breeds. The classifiers were the following: J48, SVM, Random Forest,

ISSN 2305-7254

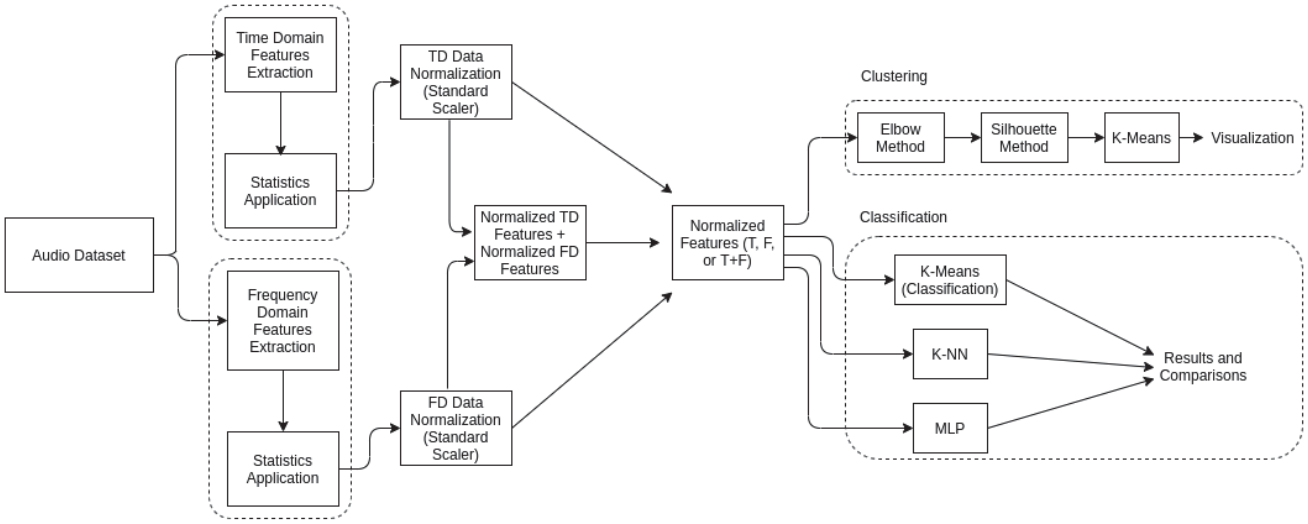

Fig. 1. Block diagram representing the work flow and the methodologies used

Bagging, Naive Bayes and Convolutional Neural Network.

There are several works in the literature concentrated on cat vocalizations. Among these is [12], which is based on a dataset produced by the authors with cat meows in various contexts from videos on YouTube and Flicker. The aim of the project was to identify the mood and context from the vocalizations of the cats. The Mel-spectrogram was extracted from each sample and given as input to a pre-trained CNN [13] trained on the Million Song dataset [14]. The aim was to use the pre-trained network as a feature extraction method. Subsequently, the authors employed different learning models on the obtained feature set and compared the results. The models used are: Random forest, k-nearest neighbor, Extremely randomized trees, Linear discriminant analysis , Quadratic discriminant analysis and Support vector machine. All classifiers were then combined to create an ensemble.

The dataset on which this work was based, i.e. CatMeows [15], was firstly used in [16]. There, the purpose was to identify three different contexts (*waiting for food*, *isolation in unfamiliar environment*, and *brushing*) from the respective vocalizations. To this end, Mel-Frequency Cepstral Coefficients (MFCC) and Temporal Modulation Features have been extracted. Lastly, different classification models were employed, i.e. Directed acyclic graphs based on Hidden Markov Models, class-specific Hidden Markov Models, Universal Hidden Markov Models, Support Vector Machine and Echo state network.

The aim of this work is to identify the cat breed with respect to an input audio, specifically the Maine Coon and European Shorthair breeds present in the audio dataset "CatMeows" [15]. To this end, different time and frequency domain features were extracted, while our focus was to capture characteristic properties of the audio structure as seen from both perspectives. Furthermore, we used statistical moments to summarize the feature information with respect to each audio sample. After normalization, we employed both unsupervised (k-means clustering) and supervised (k-NN and MLP) machine learning approaches for breed identification. We followed a thorough experimental protocol and carefully analyzed the obtained results on publicly available dataset, which conveniently enables full reproducibility of the present work. The following two sections describe the above-mentioned methodologies in terms of features and classification mechanisms. Section IV analyzes the obtained results extensively, while section V concludes this work.

II. System Overview

Figure 1 shows the proposed workflow starting from the input audio file, moving to feature extraction and normalization, ultimately leading to clustering and classification. We employed diverse features sets defined in time- and frequency-domain, which are described in the following.

A. Time Domain

As regards to the temporal domain, the methods of extraction of the features used were Amplitude Envelope (AE), Zero-Crossing Rate (ZCR) and Root-Mean-Square Energy (RMSE).

Aiming at a global representation of each audio sample, we calculated six different representative statistics, i.e. mean, median, standard deviation, max, min , and standard deviation with respect to mean. As such, for each sample and feature we obtain a vector of 6 values following the above-mentioned order. These three vectors are then concatenated (AE, ZCR, RMSE) obtaining a final vector of 18 features for the time domain.

B. Frequency Domain

As regards to the frequency domain, the extraction methods used in the following order are: Spectral Centroid (SC), Spectral Flux (SF), Spectral Rolloff (SR), and Mel-Frequency Cepstral Coefficients (MFCCs) with 13 coefficients including

the 0-th one. The process is identical to the one used in the time domain case, keeping the same statistics and order. Regarding MFCC: statistics were applied to each vector of each bin, always obtaining a vector of 6 features, concatenating the 13 vectors with respect to the index order of the bins (final vector of 13×6 = 78 features). By concatenating in the previously mentioned order, a final vector of 96 features is obtained.

Feature set fusion was implemented by concatenating the vector of 18 features of the normalized temporal domain with that of 96 of the normalized frequency domain, therefore leading to a vector of 18 + 96 = 114 features. For each of the above-described three cases, clustering is first performed, where the optimal number of clusters k was discovered via the Elbow and Silhoutte methods. Then, the identified k is considered to visualize the arrangement of the clusters (inter- and intra-cluster distances).

The accuracy of unsupervised k-means clustering was tested with $k = 2$ and evaluate the accuracy of the K-Means in predicting the ground truth (evaluating the accuracy by considering the labels directly or in reverse, and taking the best accuracy as a reference). As such, we evaluated the density of each cluster as well as the distance between samples coming from different classes in the n-dimensional space (with n denoting the number of features).

Subsequently, two learning models were employed: k-NN and MLP following an identical validation method.

During validation, the dataset is divided via holdout partitioning at the ratio 70/30 for train/test respectively; then, the training set is further divided using the ten-fold cross validation protocol. The model is trained on the sub-training set and the accuracy calculated on the remaining validation set where the model with the maximum accuracy is chosen. The confusion matrix is finally calculated on the test set, i.e. 30% of the initial dataset.

Last but not least, we constructed an MLP model representing a more refined and sophisticated machine learning-based solution. Data division and figures of metric calculation is performed in an identical way so as to achieve comparable results.

C. Feature Extraction Analysis

The specific stage comprises a fundamental process towards a successful approach, i.e. finding discriminative features with respect to the reference problem allowing the models to converge to the optimal solution. To this end, we worked within diverse domains (time and frequency), compared the achieved performances, and finally unified the obtained feature sets.

1) Time Domain features: These feature extraction methods work directly on the temporal representation of the signal, after the framing process. The considered features are the following:

- Amplitude Envelope (AE): it returns the largest (max) amplitude of the signal frame.
- Zero-Crossing Rate (ZCR): the ZCR measures the number of times the signal changes sign in a frame (i.e. the signal changes from a positive to a negative value or vice versa) divided by the length of the frame. Formally it is defined as:

$$Z(i) = \frac{1}{2K} \sum_{n=1}^{K} |sgn[x_i(n)] - sgn[x_i(n-1)]|$$

  where K is the length of the frame and $sng()$ is sign function which assigns +1 or -1.
- Root-Mean-Square Energy (RMSE): it comprises an energy metric widely used in the statistical field. It is defined as:

$$RMSE = \sqrt{\frac{1}{K} \sum_{n=1}^{K} |x(n)|^2}$$

2) Frequency Domain features: This feature extraction stage operates directly on the spectrogram obtained by mapping the original audio signal with the Fourier Transform from the time domain to that of the frequency, describing the power (magnitude) of the signal across different frequencies. In some cases more than one dimension may be returned (e.g. MFFCs). The considered methods are the following:

- Spectral Centroid (SC): it represents the center of gravity of the magnitude spectrum, usually the frequency band where most of the energy is concentrated. The value of the spectral centroid C_i of the *i*-th audio frame is defined as:

$$C_i = \frac{\sum_{k=1}^{K} kX_i(k)}{\sum_{k=1}^{K} X_i(k)}$$

- Spectral Flux (SF): it measures the spectral change between two consecutive frames and is calculated as the squared difference between normal magnitudes of the spectrum of two consecutive short-term windows:

$$Fl_{(i,i-1)} = \sum_{k=1}^{K} (EN_i(k) - EN_{i-1}(k))^2$$

  where $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{K} X_i(l)}$.
- Spectral Rolloff (SR): it is defined as the frequency below which the magnitude distribution is concentrated (around 85% -90%).
- Mel-Frequency Cepstral Coefficients (MFCCs): they comprise a cepstral representation of the signal, where the frequency bands are distributed according to the Mel-scale (a scale of frequency intervals that are perceived as equally-spaced by humans), and are very popular in the field of speech and audio processing. The process for finding these coefficients is the following: the input signal is divided into a series of overlapping frames; the magnitude spectrum of each frame is calculated; subsequently, the obtained power is mapped onto the mel-scale using overlapping triangular windows. The logarithm of each mel-frequency is then calculated and, finally, the

discrete cosine transform is applied. MFCCs are therefore the amplitudes of the mel-scaled spectrum. Typically, the first 13 MFFCs are chosen because they are considered to carry enough discriminative information in the context of various classification tasks.

D. Statistics and Normalization

For each obtained feature vector, a series of statistics to represent each available sample was extracted. The considered statistics are listed next: 1) Mean, 2) Median, 3) Standard Deviation, 4) Max, 5) Min, and 6) Standard Deviation with respect to Mean. Standard normalization techniques are applied as well, including mean removal and variance scaling.

E. Clustering Process

For each considered feature set, i.e. (time, frequency, and their combination), the following 3 steps have been carried out: 1) the first one was the execution of the Elbow Method in order to visualize the degree of error of a number of clusters ranging from 1 to 14. 2) Then follows the Silhoutte Method to identify the best k, thus choosing the k offering the highest score. 3) Finally, the distribution of points of the k clusters is visualized with a 2D graph. To this end, the feature space is reduced using Principal Component Analysis (PCA), creating a 2-dimensional dataset that is projected with the respective labels predicted by the k-Means on the original input dataset.

1) Elbow Method: The Elbow Method is a heuristic used to determine the number of clusters in a dataset. The main idea is to execute k-Means for an interval of clusters k and for each value calculate the sum of the squared distances from each point to its assigned center (distortions). The variation explained as a function of the number of clusters is then plotted in order to identify the elbow of the curve as the number of clusters to be used. In most cases, this process leads to the value of k providing the most prominent variation.

2) Silhouette Method: The silhouette method is also considered to discover the optimal number of clusters. This calculates the silhouette coefficients of each point which measure how similar a point is to its cluster compared to other clusters, providing a succinct graphical representation. The silhouette value is specifically a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). The silhouette value varies between [1, -1], where a high value indicates that the object is well suited to its cluster and poorly matched to neighboring clusters. The choice of the number of clusters generally falls on the one with the highest silhouette value.

F. Classification and Learning Models

Since the available cat breeds in the dataset are two, the problem is essentially a binary classification one. We employed models with different architectural complexity, i.e. k-NN and MLP, in order to evaluate the degree of difficulty in classifying the audio files with respect to every feature set. The validation methods are identical in order to achieve a reliable comparison. On top of these models, k-Means was used to estimate the distribution of samples in the features space.

1) k-Means: k-Means is a partition group analysis algorithm that aims to find k groups such that inter-cluster distances are maximized and intra-cluster distances are minimized. As such, the algorithm with $k = 2$ predicts the labels which are compared against the ground truth during validation. Finally, we visualized the obtained results in order to understand whether the samples are linearly separated in the features space.

2) k-Nearest Neighbors model: The k Nearest Neighbors (k-NN) model is a "lazy" learning model; to predict the label of an input sample it is based on the labels of the k closest samples (of a reference set, typically the training set) in the features space, and outputs the most frequent label among those. After early experimentations, k was set equal to 3 as it provided the best performance.

3) MLP model: The Multilayer Perceptron (MLP) comprises a feed-forward Artificial Neural Network (ANN) composed of nodes organized at different layers, i.e. an input layer, a hidden layer and an output layer. Apart from the input nodes, every other node is a neuron that uses a nonlinear activation function. Node weights are learnt via the back-propagation algorithm. MLP is able to distinguish non-linearly separable data, thus providing an alternative model to verify the linearity of the problem at hand. In this work, we adopted a network with two hidden layers: the first with a number of neurons equal to half the input size and the second layer equal to a quarter of the input size. The output layer has only one neuron since the two labels can be translated to a binary classification problem.

III. Experimental set-up

This section briefly describes the dataset and the parameterization with respect to the employed features, clustering algorithms and classification methodologies.

A. The CatMeows Dataset

The employed dataset is the "CatMeows" audio dataset [15] which consists in 440 audio files containing cat vocalizations coming from 21 different individuals belonging to 2 different breeds while recorded in 3 different contexts. Each audio is a PCM stream with a duration ranging from 1 to 3 seconds, sampled at 8000 Hz with a single channel at a bitrate of 128 kbps. The number of samples in the dataset belonging to the Maine Coon (MC) breed are 188, leaving 252 for the European Shorthair (EU) breed. Although there are more samples for the EU breed, the dataset is not highly unbalanced.

B. Feature Extraction, Statistics and Normalization

Feature extraction was carried out using librosa library [17], while, after early experimentations, the frame size was set equal to 1024 samples with a hop size of 512 samples. Moreover, the first 13 MFCCs were considered. Subsequently, the statistical functions were applied on the feature vectors of each extraction method, concatenating them in the previously mentioned order. Finally, each vector is normalized to zero mean and unit variance.

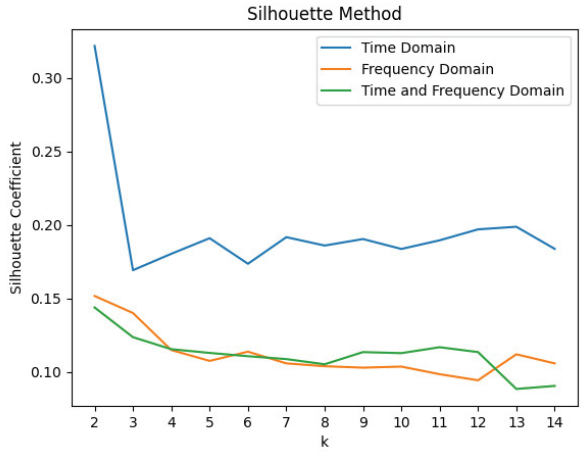

Fig. 2. Silhoutte criterion evaluating the goodness of clustering achieved using features time and frequency domain

C. Clustering

1) Silhouette and Elbow Methods: We employed two criteria to evaluate the goodness of the clustering provided by k-means, while k ranges from 1 to 14. The produced figures are shown in Fig. 2 and 3 for silhouette ($2 < k < 14$) and elbow ($1 < k < 14$) respectively.

D. Classification

1) k-means: For the prediction with k-means, we used `n_clusters = 2`, while several metrics were then calculated (both for predictions with predicted labels taken directly and inversely), i.e. confusion matrix, accuracy, etc.

2) k-NN: After early experimentations, k-NN model with $k = 3$ was used during every stage, while it was evaluated both on training and test sets. To assess the performance, we adopted the ten-fold cross validation protocol, while employing the same metrics as before for comparability.

3) MLP: The MLP model architecture is tabulated in Table I along with the design parameters allowing full reproducibility. The first hidden layer has a number of neurons equal to half of the input values rounded down (therefore for the three cases 9, 48, 57), while the second a quarter rounded to the lower integer, i.e. (4, 24, 28). Both employ the `relu` activation function. The output layer features a single neuron with `sigmoid` activation function. It should be mentioned that the learning rate was 0.001, the loss binary cross-entropy and the metric precision. The choice of the loss function is given by the fact that the prediction must be optimized for a binary problem. In addition, the following parameters were set for the reproducibility of the experiment: `np.random.seed(seed)` and `tf.random.set_seed(seed)`, where `seed = 0`.

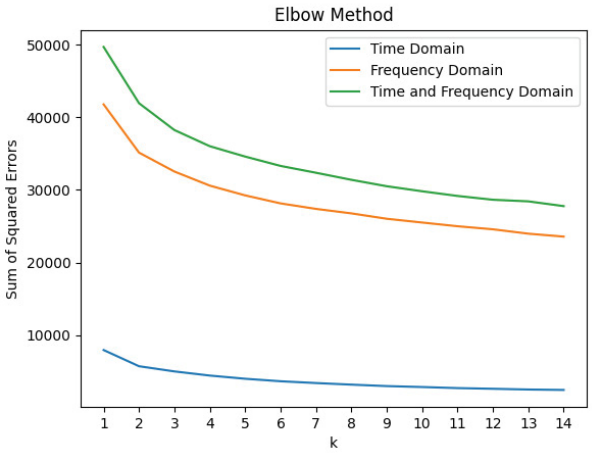

Fig. 3. Elbow criterion evaluating the goodness of clustering achieved using features time and frequency domain

TABLE I.
THE MLP ARCHITECTURE

Component	Number of neurons	Activation function
Input Layer	n (td:18, fd:96, t+f:114)	n/a
Hidden Layer 1	n/2 (td:9, fd:48, t+f:57)	ReLU
Hidden Layer 2	n/4 (td:4, fd:24, t+f:28)	ReLU
Outoput Layer	1	sigmoid
optimizer	Adam(learning_rate = 0.001)	n/a
loss	binary_crossentropy	n/a
metrics	precision	n/a

IV. EXPERIMENTAL RESULTS

This section describes the results achieved by the aforementioned clustering and classification models.

A. Clustering

Fig. 3 and 2 show the results of clustering efficacy with respect to every considered feature set. The graph of the Elbow criterion (Fig. 3) does not show a particular detachment point except perhaps slightly for $k = 2$. The result of $k = 2$ is confirmed by the Silhouette Method (Fig. 2) with a prevalent score of 0.3217 compared to the rest. It is worth noting that both criteria are characterized by relatively similar values for time-domain features, frequency-domain features as well as their combination, with the highest difference being silhoutte value for $k = 2$ in the time-domain features case.

B. Classification

1) k-Means: Table II is the confusion matrix obtained by comparing the labels predicted by k-Means with the ground truth for every considered feature set. It is interesting to note in that approximately 99% of the samples with the EU label were predicted correctly. This may be due to the fact that EU samples are well concentrated in a certain space of the time domain features. As for MC, there respective rate is 62%. Frequency domain as well as the fused set offered worse results.

TABLE II. k-MEANS CONFUSION MATRIX (%), IN THE FOLLOWING ORDER: TIME DOMAIN/FREQUENCY DOMAIN/TIME + FREQUENCY DOMAIN

Presented \ Predicted	MC	EU
MC	62 / **74** / **74**	38 / 26 / 26
EU	1.2 / 48 / 44	**98.8** / 52 / 56

2) k-NN: The results achieved by the k-NN classifier are remarkable considering the simplicity of the model, reaching 100% for MC and 96.4% when the fused feature set is employed (see Table III). It should be mentioned that the value of k providing the highest recognition rate was 3.

TABLE III. k-NN CONFUSION MATRIX (%) IN THE FOLLOWING ORDER: TIME DOMAIN/FREQUENCY DOMAIN/TIME + FREQUENCY DOMAIN

Presented \ Predicted	MC	EU
MC	88 / **100** / **100**	12 / 0 / 0
EU	1.2 / 9.6 / 3.6	**98.8** / 90.4 / 96.4

3) MLP: MLP offers results characterized by higher rates, since the specific classifiers allows processing non-linear data. Table IV tabulates the achieved rates; we observe that MLP reaches 100% for MC and 98.8% for EU when frequency domain features are used demonstrating the efficacy of the present solution. The combined use of feature sets does not seem to bring significant improvements.

TABLE IV. MLP CONFUSION MATRIX (%) IN THE FOLLOWING ORDER: TIME DOMAIN/FREQUENCY DOMAIN/TIME + FREQUENCY DOMAIN

Presented \ Predicted	MC	EU
MC	95.9 / **100** / 98	4.1 / 0 / 2
EU	7.2 / 1.2 / 0	92.8 / 98.8 / **100**

V. CONCLUSIONS

This article evaluated the performance of a wide variety of acoustic features combined with traditional machine learning algorithms to address automatic audio-based cat breed classification. The employed dataset and implementation of the experiments are publicly available[1] facilitating reproducibility. Interestingly, it was shown that the usage of statistics to discriminate the available samples was quite effective as demonstrated by the results of k-NN and the excellent performance reached by MLP. The use of different features (time and frequency) and their combination have not shown much diversity, since quite high rates were achieved only time domain features. Future developments related to the specific problem could be the use of diverse features, e.g. wavelet, combined with more advanced classification models. Finally, we intent to deploy and test the presented system by means of a smartphone application.

[1]https://github.com/williamraccagni/cat_breed_acoustic_classification

REFERENCES

[1] S. Ntalampiras and I. Potamitis, "Acoustic detection of unknown bird species and individuals," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 3, pp. 291–300, Mar. 2021. [Online]. Available: https://doi.org/10.1049/cit2.12007

[2] I. Potamitis, "Automatic classification of a taxon-rich community recorded in the wild," *PLoS ONE*, vol. 9, no. 5, p. e96936, May 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0096936

[3] S. Ntalampiras, A. Pezzuolo, S. Mattiello, M. Battini, and M. Brščić, "Automatic detection of cow/calf vocalizations in free-stall barn," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, 2020, pp. 41–45.

[4] M. P. Mcloughlin, R. Stewart, and A. G. McElligott, "Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring," *Journal of The Royal Society Interface*, vol. 16, no. 155, p. 20190225, Jun. 2019. [Online]. Available: https://doi.org/10.1098/rsif.2019.0225

[5] M. Acconcjaioco and S. Ntalampiras, "One-shot learning for acoustic identification of bird species in non-stationary environments," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 755–762.

[6] E. Prato-Previde, S. Cannas, C. Palestrini, S. Ingraffia, M. Battini, L. A. Ludovico, S. Ntalampiras, G. Presti, and S. Mattiello, "What's in a meow? a study on human classification and interpretation of domestic cat vocalizations," *Animals*, vol. 10, no. 12, p. 2390, Dec. 2020. [Online]. Available: https://doi.org/10.3390/ani10122390

[7] A. Larrañaga, C. Bielza, P. Pongrácz, T. Faragó, A. Bálint, and P. Larrañaga, "Comparing supervised learning methods for classifying sex, age, context and individual mudi dogs from barking," *Animal Cognition*, vol. 18, no. 2, pp. 405–421, Oct. 2014.

[8] S. Hantke, N. Cummins, and B. Schuller, "What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5134–5138.

[9] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, *The INTERSPEECH 2014 Computational paralinguistics challenge: cognitive physical load*, 01 2014.

[10] C. Molnár, F. Kaplan, P. Roy, F. Pachet, P. Pongrácz, A. Dóka, and Á. Miklósi, "Classification of dog barks: a machine learning approach," *Animal Cognition*, vol. 11, no. 3, pp. 389–400, Jan. 2008. [Online]. Available: https://doi.org/10.1007/s10071-007-0129-9

[11] H. Pérez-Espinosa, V. Reyes-Meza, E. Aguilar-Benitez, and Y. M. Sanzón-Rosas, "Automatic individual dog recognition based on the acoustic properties of its barks," *Journal of Intelligent Fuzzy Systems*, vol. 34, no. 5, p. 3273–3280, May 2018. [Online]. Available: https://doi.org/10.3233/JIFS-169509

[12] Y. R. Pandeya and J. Lee, "Domestic cat sound classification using transfer learning," *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS*, vol. 18, no. 2, pp. 154–160, Jun. 2018. [Online]. Available: https://doi.org/10.5391/ijfis.2018.18.2.154

[13] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," *CoRR*, vol. abs/1703.09179, 2017. [Online]. Available: http://arxiv.org/abs/1703.09179

[14] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[15] L. A. Ludovico, S. Ntalampiras, G. Presti, S. Cannas, M. Battini, and S. Mattiello, "CatMeows: A publicly-available dataset of cat vocalizations," in *MultiMedia Modeling*. Springer International Publishing, 2021, pp. 230–243.

[16] L. A. Ntalampiras, S. Ludovico, G. Presti, E. P. P. Previde, M. Battini, S. Cannas, C. Palestrini, and S. Mattiello, "Automatic classification of cat vocalizations emitted in different contexts," *Animals*, vol. 9, no. 8, p. 543, Aug. 2019. [Online]. Available: https://doi.org/10.3390/ani9080543

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.