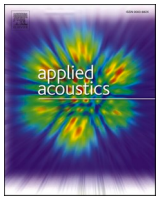
Applied Acoustics 226 (2024) 110218

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Towards Understanding Cat Vocalizations: A Novel Cat Sound Classification Model Based on Vision Transformers

Enver Kucukkulahli , Abdullah Talha Kabakus *

Department of Computer Engineering, Faculty of Engineering, Duzce University, Türkiye

A R T I C L E I N F O

Keywords:
Sound classification
Convolutional neural network
Vision transformer
Transfer learning
Deep learning

A B S T R A C T

Animal sound perception systems are highly developed compared to humans, crucial for survival in natural environments, with some species possessing specialized sensory capabilities such as vision, hearing, touch, and environmental awareness. Understanding animal sounds not only aids in their own communication and survival but also benefits humans in various fields including security, natural disaster prediction, ecological research, bioacoustics, precision agriculture, and search and rescue operations. Motivated by this fact, this study investigated the classification of cat sounds using deep learning models based on Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures. Cat vocalizations, represented as mel-spectrograms, were classified using models trained on a diverse dataset of cat sounds. Experimental results demonstrated the superiority of the proposed model based on *Microsoft*'s *BERT Pre-Training of Image Transformers* (*BEiT*) over the state-of-the-art as it obtained an exceptional accuracy of 96.95%. Additionally, it was observed that the proposed models based on ViT outperformed CNN-based models, highlighting the efficacy of transformer architectures in capturing complex patterns within audio data. These findings underscore the potential of ViT architectures in decoding animal communication systems and advancing wildlife conservation efforts.

1. Introduction

Animals have advanced perception systems compared to human beings due to the necessity of helping them survive in their natural environments. Some animals have special sensory capabilities, such as vision, sight, feeling, and awareness of natural changes [1]. The sounds of animals are not only helpful for themselves but also for human beings in areas such as security, prediction of natural disasters, ecological research, bioacoustics, precision agriculture, and search and rescue operations. In addition to this, understanding animal sounds would let human beings understand their intentions well.

Pet animals (*a.k.a.* pets) have always been great friends of human beings for numerous reasons, such as (*i*) their loyalty and affection, (*ii*) increasing emotional well-being (e.g., they reduce loneliness), (*iii*) physical health benefits, (*iv*) stress reduction, (*v*) being companions to physical activities, (*vi*) social interaction, and (*vii*) security. Cats are one of the most, if not the most, loved pets in the world. The beginning of this great relationship is still uncertain, but it is generally assumed that the cat was first domesticated in Egypt during the second millennium BC, but a fossil record was recently found in Cyprus and was dated to around 7,000 years ago [2]. It is argued that the relationship between cats and humans seems unique compared to other forms of human-animal interaction, as the behavioral patterns directed towards people closely resemble those observed in social interactions among cats [3]. As per a recent report [4], 66% of households in the United States, equivalent to 86.9 million homes, are pet owners, and 46.5 million households specifically own a cat. As presented in Fig. 1, the *Google Trends* index of the search term "*cat*" [5] has shown an upward trend, rising from 75 to 95 over the past decade. This statistic indicates a growing interest in cats.

During the past decade, researchers began exploring the promise of visually representing sound signals to benefit from powerful image classifiers, which are mostly based on Deep Neural Networks (e.g., Convolutional Neural Networks or CNNs) as they have provided superior results compared to other classifiers when it comes to image classification [6]. Cat sounds represent the most distinctive intra-specific vocalization among felids, whether wild or domesticated [7]. Motivated by this fact and recognizing the significant popularity of cats, this study aims to introduce an innovative and robust cat sound classifier. This classifier is specifically designed to overcome common challenges such as biodiversity, species variation, and age differences. The objectives of

* Corresponding author.
E-mail address: talhakabakus@duzce.edu.tr (A.T. Kabakus).

https://doi.org/10.1016/j.apacoust.2024.110218
Received 28 March 2024; Received in revised form 4 July 2024; Accepted 3 August 2024
Available online 7 August 2024

this study are as follows: (*i*) investigate the classification of cat sounds, (*ii*) evaluate the performance of ViT and CNN models, (*iii*) utilize mel-spectrograms for sound representation, (*iv*) achieve high classification accuracy, (*v*) explore the efficacy of transformer architectures, (*vi*) contribute to animal sound perception systems, and (*vii*) advance wildlife conservation efforts. Consequently, the utilization of a large and diverse dataset that accurately represents these factors becomes imperative. All Machine Learning (ML) models learn knowledge from provided data, adhering to the common adage "garbage in – garbage out," emphasizing the significance of the quality of input data in determining the model's output. In addition, the ML model itself should be thoroughly optimized and possess sufficient complexity to effectively capture and learn the intricate patterns present in the dataset. Transfer learning is a technique that allows the utilization of pre-trained, state-of-the-art ML models for distinct yet related tasks. Deep Learning (DL) platforms (e.g., *TensorFlow* [8], *PyTorch* [9]) already provide some of the state-of-the-art DL models. After visually representing cat sounds, we employed several state-of-the-art DL models through transfer learning. The main contributions of this study are listed as follows:

- *Visual representation of sound signals.* The cat sounds were represented as images to benefit from state-of-the-art image classifiers.
- *A novel cat sound classification model based on a wide range of state-of-the-art image classifiers.* The architecture of the proposed model is finalized after conducting experiments on a wide range of state-of-the-art image classifiers based on DL. Unlike the related work, we also employed Vision Transformer (ViT) models.
- *Extensive hyperparameter optimization.* The hyperparameters of the proposed model were optimized through an automated, extensive hyperparameter optimization task. This process makes the obtained classifier unique as its hyperparameters are uniquely suited to the problem at hand.
- *Extensive sound visualization experiments.* Unlike the related works that directly opted for a sound visualization technique, various visual representations of sound signals were evaluated, and the one that provided the best classification accuracy was chosen.
- *Highly accurate, yet fast classifier:* The proposed classifier obtained an accuracy as high as 96.95% and an inference time as low as 2.841 milliseconds, which are both critical for applications such as real-time wildlife monitoring, pet behavior analysis, and interactive systems involving animals.

The remainder of the paper is structured as follows: Section 2 presents an overview of the related work, highlighting key studies relevant to the topic. Section 3 delves into the materials and methods employed for this study, providing a detailed description of the dataset construction and the proposed model. Section 4 is dedicated to presenting the experimental results and discussion, where the findings from the conducted experiments are analyzed and interpreted. Finally, Section 6 serves as the conclusion of the paper, summarizing the key findings and outlining potential directions for future research endeavors.

2. Related work

Traditional rule-based approaches have been employed in the related work. Li and *Wu* [27] proposed a method for recognizing animal sounds in various noisy environments with different Signal-to-Noise Ratios (SNRs). Recognizing animal sounds automatically in real-world settings aids in monitoring and protecting animal populations by analyzing their habits and distributions. Addressing the challenge of maintaining recognition accuracy under low SNR conditions, the authors introduce a double-feature approach combining projection feature and Local Binary Pattern Variance (LBPV) feature, used alongside random forests for classification. The projection feature is derived from spectrogram projections, while the LPBV feature is obtained by accumulating variances for uniform local binary patterns in the spectrogram. Experimental results demonstrate that the proposed method can recognize a wide range of animal sounds, achieving over 80% recognition accuracy even under 10*dB* SNR conditions, showcasing its robustness in noisy environments.

ML-based methods have also been employed in the related work. *Mohmmad et al.* [28] addressed the emerging trend of recognizing animals in forests using sound classification, focusing on classifying tiger roar sounds with an optimized CNN model. Sounds were extracted from *YouTube* and augmented using Noise Injection and Pitch Shift techniques, creating a dataset of 1,200 audio samples from four animal species. Features such as MFCC, ZCR, and Mel-Spectrogram were used to enhance performance. The model's interpretability ensures practical applications in wildlife conservation and monitoring, aiding forest department personnel. The objective was to accurately identify animal vocalizations, specifically from tigers, leopards, elephants, and otters. Among several classifiers tested, including SVM, Random Forest, LSTM, and CNN, the optimized CNN achieved the highest accuracy rate of 91%, demonstrating its effectiveness for this task. *Vithakshana* and *Samankula* [29] discuss the challenges and solutions for classifying animals within the kingdom "Animalia" using acoustical methods when visual methods are impractical. The study emphasizes the role of classification systems in bioacoustics monitoring, which is crucial for fields such as animal science, zoology, and environmental studies. The research introduces an IoT-based acoustic classification system using CNN to monitor ecosystems. The system collects audio data, preprocesses it using MFCC, and employs a CNN architecture for training. Using 400 sound clips of 10 animal species, the network was trained with various gradient descent optimizers, achieving the highest accuracy of 91.3% with *AdaDelta*, *Gradient Descent*, and *RMSProp*, with *AdaDelta* showing the most stable learning. The study suggests future work to include larger datasets to further improve accuracy. *Sharma et al.* [30] presented a comprehensive study on the classification of animal vocalizations for estrus identification using machine learning and deep learning models. The research focused on buffalo vocalizations, utilizing MFCC for feature extraction to capture the unique frequency and temporal patterns of vocalizations. Various models, including SVM, Naive Bayes, Random Forest, kNN, CNN, Recurrent Neural Networks (RNN), Convolutional Recurrent Neural Networks (CRNN), and ResNet34, were trained and evaluated. The CNN and RNN models excelled in identifying estrus vocalizations,

Fig. 1. The *Google Trends* index of the search term "*cat*" has shown an upward trend, rising from 75 to 95 over the past decade.

achieving high accuracy, while the CRNN, kNN, and ResNet34 models showed lower performance. This study highlights the effectiveness of automated estrus identification in wildlife monitoring, demonstrating that machine learning and deep learning can significantly enhance conservation efforts by providing efficient and scalable solutions. Future research could explore transfer learning, ensemble modeling, and broader datasets to further improve these techniques for wildlife monitoring and conservation.

Hybrid models have also been employed in the related work. *Ko et al.* [23] addressed the challenge of insufficient discernibility in training databases for pattern classification, particularly with similar animal sounds. The authors proposed a novel approach combining multiple pre-trained CNNs to generate mid-level features for each class, merged into a unified CNN unit with SVM for the overall classification. Using an animal sound database with 3 classes and 102 species, the method outperformed conventional techniques. This innovative approach enhances classification accuracy but may be complex and dataset-specific, suggesting further research for broader applicability and optimization. *Lu et al.* [26] addressed the challenge of environmental sound classification (ESC) by proposing a novel CNN model leveraging transfer learning. The approach represents sound as an RGB image, with each channel corresponding to specific audio features such as the Log-Mel spectrogram, scalogram, and MFCC. The CNN architecture, based on the Xception model known for its performance on the JFT dataset, was trained using this representation. The proposed CNN obtained an accuracy of 75.3% on the test set. Test results demonstrated that the proposed approach achieves improved ESC accuracy compared to existing methods. This innovative approach holds promise for enhancing the accuracy of ESC tasks, offering potential applications in various domains including environmental monitoring and conservation.

The majority of the related works have employed DL. Animal sound classification is a vibrant research area. Research in the area of animal sound classification can be broadly classified into two categories: (*i*) approaches based on CNN and (*ii*) fingerprinting [10], which entails creating a concise audio representation to compare audio segments in relation to their similarity and dissimilarity [11]. In [12], fusions of CNNs combined with engineered features were employed for fish identification. In [13], DL was combined with shallow learning for bioacoustics bird species classification. In [14], the authors demonstrated a combination of DL with fingerprinting by proposing a Siamese Neural Network (SNN) that produced descriptions of sound signals. In [15], an approach based on dissimilarity space learning, where a distance model was obtained by training an SNN on dissimilarity values was proposed. In [16], clustering methods were employed to convert the spectrograms of the cat dataset into a collection of centroids. These centroids were then utilized to create a vector space representation for each pattern. Subsequently, the generated vectors were fed into a wide range of traditional ML models, namely, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (kNN), and Latent Dirichlet Allocation (LDA) as well as an ensemble model that averages the employed ML models. According to their experimental results, the proposed ensemble model obtained the best accuracy, an accuracy of 87.76% among all classifiers. In [1], Frequency Division Average Pooling (FDPA) was employed as the feature extractor and CNN and Convolutional Deep Belief Network (CDBN) were employed as the classifiers. According to their experimental results, the best accuracy, an accuracy of 91.13%, was obtained when the proposed CDBN was employed as the classifier. In [17], the authors constructed their own cat sound dataset, which consisted of 478 samples. They employed two sets of acoustic parameters, namely, (*i*) mel-frequency cepstral coefficients and (*ii*) temporal modulation features for capturing the emission context. Then, directed acyclic graphs-Hidden Markov Models dividing the problem space were employed to model these acoustic parameters. One major limitation of this study is that the dataset is tiny in terms of the number of samples (478) and target classes (3) it contains. It is worth mentioning that they also employed some other classification models such as class-specific Hidden Markov Models, universal Hidden Markov Models, SVM, and echo state networks, but the best classification accuracy, an accuracy of 95.94%, was obtained when directed acyclic graphs-Hidden Markov Models were employed. *Ntalampiras et al.* [18] proposed a cat sound classification approach based on *YAMNet*, a publicly available pre-trained Deep Neural Network (DNN). The proposed classifier operated on Mel-scaled spectrograms. Unlike most of the related works, they used their own dataset that was collected from several online resources, including but not limited to, *YouTube* and *freesound*. According to their experimental results, the proposed model obtained a False Positive (FP) index of 5.7% and a False Negative (FN) index of 9.1%. *Purrai* [19] is a DNN, a modified version of *Google*'s *Vggish*, that was proposed to interpret domestic cat language. According to their experimental results, *Purrai* obtained a top-1 accuracy of 74.1%. The major drawback of this study is the lack of hyperparameter optimization. In [20], the authors proposed ensembles of classifiers based on DNN for the cat and bird sound classification. Unlike the other works, the authors employed data augmentation techniques, which boosted the classification accuracy per the conducted experiments. According to the experimental results, the proposed model obtained an accuracy of 91.7%. In [21], the authors proposed an approach based on the combination of three DL techniques, namely, (*i*) *You-Only-Look-Once* (*YOLO*), (*ii*) *Single Shot Detection* (*SSD*), and (*iii*) *InceptionV3* for the classification of cat pain. According to the experiments conducted on a dataset collected from veterinary practices, the proposed model obtained an accuracy of 93%. Unlike the rest of the reviewed works, in [22], the authors proposed an approach based on manually annotated geometric landmarks and CNN, namely, *ResNet50*, for the recognition of pain in cats from cat facial images instead of cat vocalizations. To this end, they constructed their own dataset of 84 client-owned cats. Cats underwent scoring by veterinary experts along with thorough clinical history documentation. These scores were subsequently utilized for training AI models through two distinct approaches. According to their experimental results, the landmark-based model outperformed the DL model based on *ResNet50*, achieving an accuracy exceeding 77% in pain detection. *Sasmaz* and *Tek* [24] explored animal sound classification using deep learning, proposing a system based on CNNs. Sound files were preprocessed to extract Mel Frequency Cepstral Coefficients (MFCC) using the *librosa* library. The dataset comprised 875 animal sound samples from 10 different animal types. The study reported classification confusion matrices and results from various gradient descent optimizers, achieving the highest accuracy of 75% with *Nesterov-accelerated Adaptive Moment Estimation* (*Nadam*). While the method shows promise, the accuracy indicates room for improvement and highlights the challenge of animal sound classification. *Weninger* and *Schuller* [25] investigated data-based recognition of animal sounds using a real-world database from the *Humboldt-University Animal Sound Archive*. The classifiers were challenged to discriminate between species without preselecting favorable cases, considering variations in age and stance. The authors defined classification tasks to aid information retrieval and indexing of large sound archives. They compared dynamic and static classification methods, including Hidden Markov Models, recurrent neural networks with Long Short-Term Memory (LSTM), and Support Vector Machines (SVM), using various features typical in sound classification and speech recognition. The study achieved up to 81.3% accuracy on a 2-class task and 64% on a 5-class task, highlighting the potential and challenges of the approach. *E2E-ResNet* [31] is an end-to-end *ResNet* model designed to synthesize speech signals from the silent video of a speaking individual. The model uses a convolutional encoder-decoder framework to convert video frames into latent visual features, which are then decoded into spectrograms and converted to speech waveforms. Experimental results on the *GRID* and *TCD-TIMIT* databases demonstrate that the *E2E-ResNet* model produces realistic and intelligible speech, outperforming competing approaches with improvements of 3.077% in speech quality and 2.593% in speech intelligibility.

Table 1 lists a comparison of the related work in terms of the target

Table 1
A comparison of the related work.

Related Work	Target Animal(s)	Employed Technique(s)	Limitation(s)	Classification Accuracy
[16]	cat	Traditional ML models	Lack of employment of more complex models and low accuracy	87.76%
[1]	cat	FDAP as feature extractor and CDBN as classifier	Low accuracy	91.13%
[32]	cat and bird	DNN	Low accuracy for cats	88.47%(cat) and 96.03% (bird)
[17]	cat	Directed acyclic graphs-Hidden Markov Models	Limited dataset in terms of number of samples (478) and number of target classes (3)	95.94%
[18]	cat	DNN	–	FP: 5.7%, FN:9.1%
[19]	cat	DNN	Lack of hyperparameter optimization and low accuracy	74.1%
[20]	cat and bird	DNN with data augmentation	Low accuracy for cats	91.7%(cat) and 96.8% (bird)
[21]	cat	DNN	–	93%
The proposed work	**cat**	**ViT**	–	95.96%

animal, employed technique(s), limitation(s), and obtained classification accuracy.

3. Material and method

In this section, we provide detailed descriptions of the dataset construction and the proposed model in the subsections.

3.1. Dataset construction

A gold standard dataset is a key necessity for the success of any ML model. To this end, we use *CatSound* [16], a dataset comprised of 5,922 cat sound recordings from 10 classes as follows: (*i*) *warning*, (*ii*) *angry*, (*iii*) *defence*, (*iv*) *fighting*, (*v*) *happy*, (*vi*) *hunting mind*, (*vii*) *mating*, (*viii*) *mother call*, (*ix*) *paining*, and (*x*) *resting*. The dataset can be considered as balanced as it contains approximately 600 sound recordings for each class. The sound recordings of this dataset were collected from online sources such as *YouTube* and *Flickr*. The distribution of cat sound classes in the *CatSound* dataset with the number of samples within each class is presented in Fig. 2.

The cat sound recordings (*.mp*3 files) were transformed into images (*.png* files) through *librosa* [33], an open-source Python library for audio and music analysis. While it is possible to output Mel-spectrograms using custom code written by the authors, we utilized *librosa* for (*i*) efficiency and reliability and (*ii*) standardization, the reproducibility of our results, and allows other researchers to easily replicate our methodology. Throughout this transformation process, three visual representations of sounds, namely, (*i*) waveform, (*ii*) spectrogram, and (*iii*) mel-spectrogram were evaluated to reveal the one that provided the highest classification accuracy. Eventually, the constructed dataset comprised 5,922 images from the aforementioned 10 classes in the shape of $(310, 308, 3)$. The dataset construction process of the proposed study is illustrated in Fig. 3.

3.2. Proposed model

To identify the optimal classifier for accurately classifying cat sounds, we conducted a thorough evaluation of a diverse range of deep learning models spanning various architectures. In the subsections, we detail the proposed models.

3.2.1. Proposed model based on CNN

The mathematical representation of a CNN is given as follows: A CNN processes the input $X \in \mathbb{R}^{H \times W}$ through a series of layers. Each convolutional layer applies a set of K learnable filters $W_k \in \mathbb{R}^{F \times F}$ (where F is the filter size), producing feature maps $Y_k = \sigma(X * W_k + b_k)$, where * denotes the convolution operation, b_k is the bias, and σ is a non-linear activation function. These feature maps are typically followed by pooling layers, which downsample the spatial dimensions. After several convolutional and pooling layers, the output is flattened into a vector and fed into fully connected (FC) layers, represented as $z = \sigma(W_{fc}x + b_{fc})$, where W_{fc} and b_{fc} are the weights and biases of the FC layer, respectively. The final layer is a Softmax layer, producing class probabilities $y = softmax(W_{out}z + b_{out})$, $W_{out} \in \mathbb{R}^{N \times C}$ and $b_{out} \in \mathbb{R}^{C}$ are the weights and

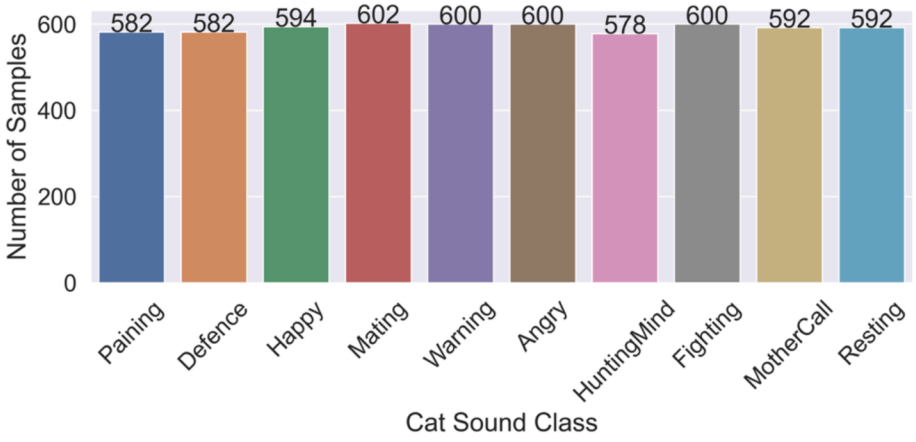


Fig. 2. The distribution of cat sound classes in the *CatSound* dataset with the number of samples within each class.

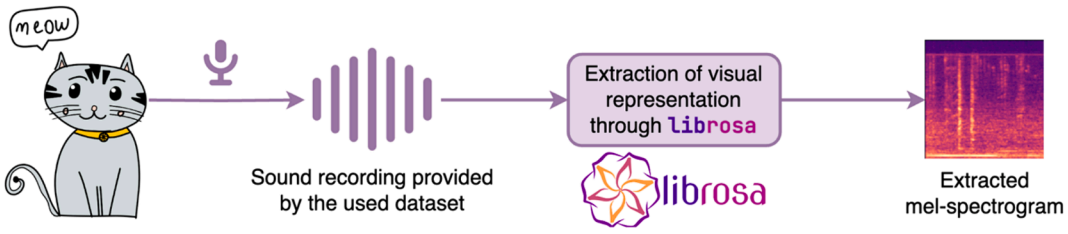


Fig. 3. An illustration of the dataset construction process of the proposed study.

biases for the C classes.

Considering the tremendous success of DL in many classification problems, we aimed to propose a model based on DL. Transfer learning is a technique to transfer the knowledge gained by an existing model to another, but similar task. When the dataset that will be used to train an ML model is limited in terms of the number of samples it has, transfer learning is the primary step to be taken while constructing an effective model. Furthermore, transfer learning contributes to reducing the necessary training time by avoiding training the entire network from the ground up. Additionally, it enhances the model's generalization capability by leveraging knowledge acquired from the source task, thereby encompassing a broader understanding of features present in the data. To this end, we employed a wide range of state-of-the-art pre-trained DL models, namely, (*i*) *Xception*, (*ii*) *InceptionV3*, (*iii*) *InceptionResNetV2*, (*iv*) *DenseNet121,* (*v*) *DenseNet169*, (*vi*) *DenseNet201*, (*vii*) *VGG19*, (*viii*) *MobileNetV2*, and (*ix*) *ResNet50V2*, through transfer learning. These pre-trained models are briefly described as follows: *Xception* is a deep CNN architecture that is known for its depth-wise separable convolutions, which help reduce the number of parameters while maintaining expressive power. It is particularly suitable for tasks with limited computational resources. *InceptionV3* is part of the *Inception* family of architectures, known for its inception modules that use multiple filter sizes in parallel to capture features at different scales. It strikes a balance between model size and performance. *InceptionResNetV2* combines the principles of the Inception architecture with residual connections from *ResNet*. This results in a deep network with improved performance and better gradient flow during training. *DenseNet* architectures (*DenseNet121*, *DenseNet169*, and *DenseNet201*) introduce dense connectivity patterns between layers, where each layer receives direct input from all preceding layers. This promotes feature reuse and gradient flow, leading to improved parameter efficiency and performance. *VGG19* is a CNN architecture with a simple and uniform structure, consisting of 19 layers with small (3×3) convolutional filters. It is known for its simplicity and effectiveness, making it a popular choice for baseline comparisons. *MobileNetV2* is designed for mobile and embedded vision applications, with a focus on efficiency and speed. It utilizes depth-wise separable convolutions and linear bottlenecks to reduce computational cost while maintaining performance. *ResNet50V2* is part of the *ResNet* (Residual Network) family of architectures, known for its residual connections that the enable training of very deep networks. *ResNet50V2* specifically has 50 layers and includes various improvements over the original *ResNet* architecture. A comparison of the employed pre-trained CNNs in terms of release year, number of parameters, estimated multiply-accumulate counts (MACs), depth (number of layers), top-1 accuracy (the proportion of predictions where the model's most confident class prediction matches the ground truth label) obtained on the *ImageNet* dataset, and key feature(s) is given in Table 2. The proposed CNN-based models were implemented using *Keras* [34], which is a high-level API for the implementation of DNNs using several backends. In this study, *TensorFlow*, which is *Google*'s widely used ML suite and the default backend of *Keras*, was opted as the backend.

During the transfer learning process, we utilized the pre-trained models as feature extractors, thereby freezing the weights of their layers. The layers responsible for classification were omitted from the constructed models based on these pre-trained models. Subsequently, a Dense layer — comprising fully connected neurons, where each neuron is connected to every neuron in the preceding layer—was sequentially added. This Dense layer consisted of as many units as the number of target classes, which in this case was 10 for the classification of cat sounds. Since the cat sound classification is a multiclass classification problem, the *Softmax* was employed as the activation function of this Dense layer. The *Glorot* (*a.k.a. Xavier*) *Uniform* [35] was employed as the kernel initializer, and the bias vector was initialized with zeros. *Categorical Cross-Entropy* was employed as the loss function for the constructed model, given that the handled problem involves multiclass classification. The architecture of the proposed models based on pre-trained CNNs is illustrated in Fig. 4.

3.2.2. Proposed model based on vision transformer

In addition to employing state-of-the-art CNNs through transfer learning, we also employed state-of-the-art ViTs, which are transformer-based neural network architectures designed for image classification tasks. Even though CNNs have been the dominant architecture for computer vision tasks, ViT introduces a transformer architecture, which was originally developed for Natural Language Processing (NLP), into the domain of computer vision. A ViT comprises three key components, namely (*i*) patch embeddings, which are obtained by dividing the input image into fixed-size patches and serve as the input tokens for the transformer encoder, (*ii*) a transformer encoder, which is responsible for processing the patch embeddings through self-attention mechanisms,

Table 2
A comparison of the employed pre-trained CNNs.

Model	Release Year	Number of Parameters (Millions)	MACs (Millions)	Depth	Top-1 Accuracy (%)	Key Feature(s)
Xception	2017	22.9	45.8	71	79	Depth-wise separable convolutions
InceptionV3	2015	23.8	47.8	159	77.9	Inception modules
InceptionResNetV2	2016	55.8	111.8	572	80.3	Inception modules and residual connections
DenseNet121	2016	8.1	16.2	121	74.6	Dense connectivity between layers
DenseNet169	2016	14.3	28.6	169	76.2	Dense connectivity between layers
DenseNet201	2016	20.2	40.4	201	77.3	Dense connectivity between layers
VGG19	2014	143.6	287.4	26	71.3	Uniform architecture
MobileNetV2	2018	3.4	7	53	71.3	Depth-wise separable convolutions and inverted residuals
ResNet50V2	2016	25.6	51.2	50	77.2	Residual connections

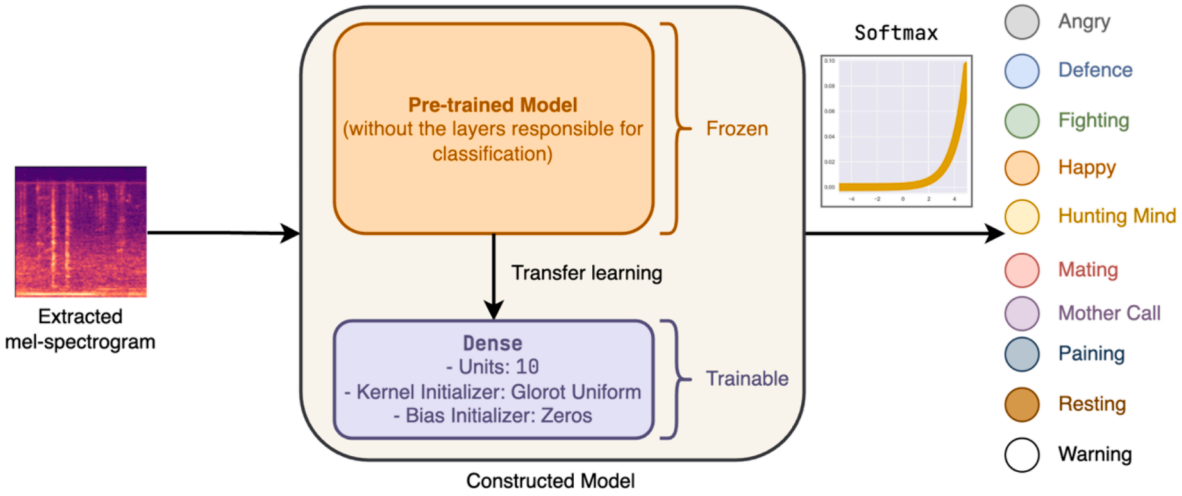


Fig. 4. An illustration of the architecture of the proposed models based on pre-trained CNNs.

feed-forward neural networks, normalization layers, residual connections, and positional encodings, and (*iii*) a classification head, which is responsible for predicting the class label of the given input image. An overview of the architecture of ViT is presented in Fig. 5. The mathematical representation of a ViT is given as follows: ViT processes the input $X \in \mathbb{R}^{H \times W}$ by first dividing it into N non-overlapping patches of size $P \times P$. Each patch is flattened into a vector $x_i \in \mathbb{R}^{P^2}$ and linearly projected to a D-dimensional embedding $z_i = W_p X_i + b_p$, where $W_p \in \mathbb{R}^{D \times P^2}$ and $b_p \in \mathbb{R}^D$. Position embeddings $E_{pos} \in \mathbb{R}^{N \times D}$ are added to these embeddings, resulting in $z_i' = z_i + E_{pos,i}$. The sequence of embedded patches $Z = [z_1', z_2', \cdots, z_N']$ is fed into a Transformer encoder, where each layer applies a multi-head self-attention mechanism, defined as $Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, followed by a position-wise Feed-Forward Network (FFN) $FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$. A special classification token z_0 is prepended to the sequence, and the output corresponding to this token is passed through a linear layer to produce the final classification logits $y = softmax(W_c z_0 + b_c)$, where $W_c \in \mathbb{R}^{D \times C}$ and $b_c \in \mathbb{R}^C$ are learnable parameters, with C being the number of classes.

We covered two widely used ViTs, namely, (*i*) *Google ViT* [36] and (*ii*) *Microsoft's BERT Pre-Training of Image Transformers* (*BEiT*) [37]. *Google ViT* is a transformer encoder model initially pre-trained on a substantial collection of images known as *ImageNet-21 k*, featuring a resolution of 224×224 pixels. Following this pre-training phase, the model undergoes fine-tuning on *ImageNet* [38], a dataset comprising 1 million images distributed across $1,000$ classes, also at a resolution of 224×224. Through pre-training, the model acquires an intrinsic understanding of images, enabling it to extract features beneficial for subsequent tasks. During preprocessing, images are resized to a uniform resolution of 224×224 pixels and normalized across the RGB channels using a mean of $(0.5, 0.5, 0.5)$ and a standard deviation of $(0.5, 0.5, 0.5)$. *Microsoft BEiT* is also a ViT model that was pre-trained on the *ImageNet-21 k* dataset and fine-tuned on *ImageNet*. *BEiT*'s pre-training objective involves predicting visual tokens from the encoder of *OpenAI*'s *DALL-E*'s *VQ-VAE*, utilizing masked patches as the basis for prediction. During the employed preprocessing, the images are resized to 224×224 dimensions and then normalized across the RGB channels. This normalization involves adjusting the mean to (0.5, 0.5, 0.5) and the standard deviation to (0.5, 0.5, 0.5). The model was presented with images in the form of a sequence of fixed-size patches, each with a resolution of $16 \times$

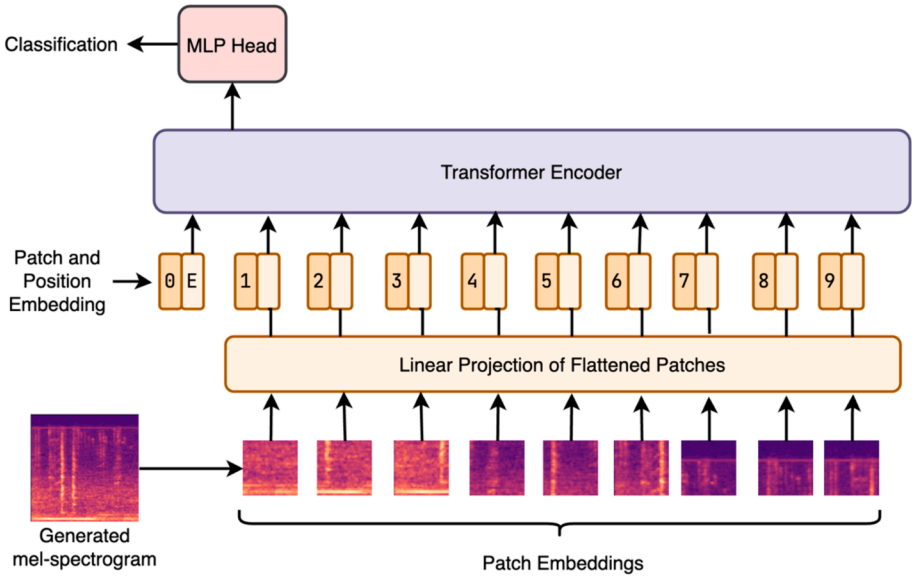


Fig. 5. An overview of the architecture of ViT, where E represents extra-learnable class embeddings.

16 pixels. These patches were then linearly embedded for processing. A comparison of the employed ViTs in terms of release year, architecture, attention mechanism, input representation, pre-training approach, training data, number of parameters, and estimated multiply-accumulate operation (MAC) count, is given in Table 3. Both models are available at the *Hugging Face* [39] platform, which is an open-source community focused on NLP and Artificial Intelligence (AI). *Hugging Face* is best known for its *Transformers* [40] library, which provides a developer-friendly interface for employing state-of-the-art transformer-based models. We employed both ViTs through this library.

3.3. Hyperparameter optimization

Hyperparameters are the parameters of DL models that are neither learned from the training data nor have formulas but are set empirically. Optimization of the hyperparameters plays a key role in enhancing the learning ability of any ML/DL model. The hyperparameters of a typical DL model include, but are not limited to, learning rate, batch size, and optimization algorithm. Hyperparameters are optimized by employing several techniques such as Grid Search, Random Search, Bayesian Optimization, Genetic Algorithms, and Hyperband [41]. We employed Hyperband for the optimization of the hyperparameters of the proposed model. Within the scope of the employed hyperparameter optimization task, we covered five hyperparameters, namely, (*i*) learning rate, which determines the size of the steps taken during optimization (e.g., gradient descent) to update the model parameters, (*ii*) optimization algorithm, which plays a crucial role in training DL models by adjusting the model's parameters to minimize a chosen loss function, (*iii*) batch size, which refers to the number of training examples utilized in one iteration during the optimization process, (*iv*) pooling type, namely, (1) Average Pooling, which applies pooling through the average function and (2) Max Pooling, which applies pooling through the maximum function, (*v*) test set ratio, and (*vi*) base model. Regarding the optimization algorithms, we covered five widely used algorithms, namely, (*i*) *Adaptive Moment Estimation* (*Adam*), (*ii*) *Stochastic Gradient Descent* (*SGD*), (*iii*) *Root Mean Squared Propagation* (*RMSprop*), (*iv*) *Adadelta*, and (*v*) *Adamax*. For both learning rate and batch size, commonly used values were evaluated. The loss calculated for the validation set was set as the objective of the employed optimization. Table 4 presents the optimized hyperparameters of the proposed model alongside their evaluation values, with the obtained optimized values given in bold.

4. Experimental results and discussion

In this section, we present the experimental results and discussion in the light of conducted experiments.

4.1. Comparison of sound visualization techniques

Three sound visualization techniques, namely, (*i*) waveform, (*ii*) spectrogram, and (*iii*) mel-spectrogram were covered for the proposed study to reveal which one provides the best classification accuracy. For this experiment, the images generated by each of these techniques were yielded into a wide range of classifiers constructed using the same architecture provided by Fig. 4. Based on the experimental results depicted in Fig. 6, it is evident that the mel-spectrogram yielded better classification accuracy compared to the waveform and spectrogram across all employed classifiers. Consequently, we chose to represent the cat sounds visually using the mel-spectrogram. Specifically, we generated 128 Mel bands for this purpose. The audio time series of the sound recordings with their sampling rates were preserved. The Fast Fourier Transform (FFT) window length and the number of samples between consecutive frames were configured to 2,048 and 512, respectively. These values were the default settings predefined by *librosa*. The *Hanning* (*Hann*) function was employed as the window function.

Table 3
A comparison of the employed ViTs.

Feature	Google ViT	Microsoft BEiT
Release year	2021	2022
Architecture	Transformer-based	Transformer-based with bottleneck mechanism
Attention mechanism	Self-attention	Bottleneck attention
Input representation	Pixel patches from images	Pre-extracted image features
Pre-training approach	Token-based	Image feature-based
Training data	*ImageNet*	*ImageNet*
Number of parameters	125 million	2 billion
Estimated multiply-accumulate operation (MAC) count	250 million	4 billion

Table 4
The optimized hyperparameters of the proposed model with their evaluation values. The obtained optimized values are given in bold.

Hyperparameter	Evaluated Values
Learning rate	$1 \times e^{-2}$, $1 \times \mathbf{e^{-3}}$, $1 \times e^{-4}$
Optimization algorithm	*Adam*, *SGD*, *RMSprop*, *Adadelta*, ***Adamax***
Batch size	8, 16, 32, 64
Pooling type	**Average Pooling**, Max Pooling
Test set ratio	0.4, 0.3, 0.2, 0.1
Base model	• *Xception* • *InceptionV3* • *InceptionResNetV2* • *DenseNet121* • *DenseNet169* • *DenseNet201* • *VGG19* • *MobileNetV2* • *ResNet50V2* • ***Google ViT*** • *Microsoft BEiT*

4.2. Evaluation of proposed models

All the proposed models were trained for 30 epochs under the same training configuration to obtain a fair comparison. The hyperparameters of the training were finalized through the experimental results of the employed hyperparameter optimization task, which is described in the previous section. In light of this task, *Adam* was employed as the optimization algorithm with a learning rate of $1 \times e^{-3}$. A batch size of 8 was utilized. Average Pooling was employed to reduce dimensionality and noise, prevent overfitting, and ensure translation invariance. In adherence to common practices in training DL models, the entire dataset was divided into three subsets: (*i*) training set, (*ii*) validation set, and (*iii*) test set. Initially, 20% of the whole dataset was allocated as the test set. The remaining data was further partitioned into two sets: 20% of this remaining data was designated as the validation set, while the remaining portion was utilized for training the proposed models. All proposed models underwent training and validation processes on the designated training and validation sets, respectively.

In a similar fashion, the proposed models were evaluated on the same test set. Fig. 7 presents the accuracy obtained by the proposed models on the test set. According to the experimental results, the proposed model based on *Microsoft BEiT* obtained the highest accuracy, an accuracy of 95.96%. The proposed model based on *Google ViT* followed *Microsoft BEiT* by obtaining an accuracy of 94.11%. Another noteworthy experimental result is that the proposed models utilizing ViT obtained higher accuracy compared to models based on CNN. Table 5 provides a comparative analysis of related studies, emphasizing the classification accuracy attained. It is important to note that this table exclusively includes related studies employing the same dataset as the proposed study to ensure a fair comparison. As evident from this table, the proposed model based on *Microsoft BEiT* surpassed the performance of the related

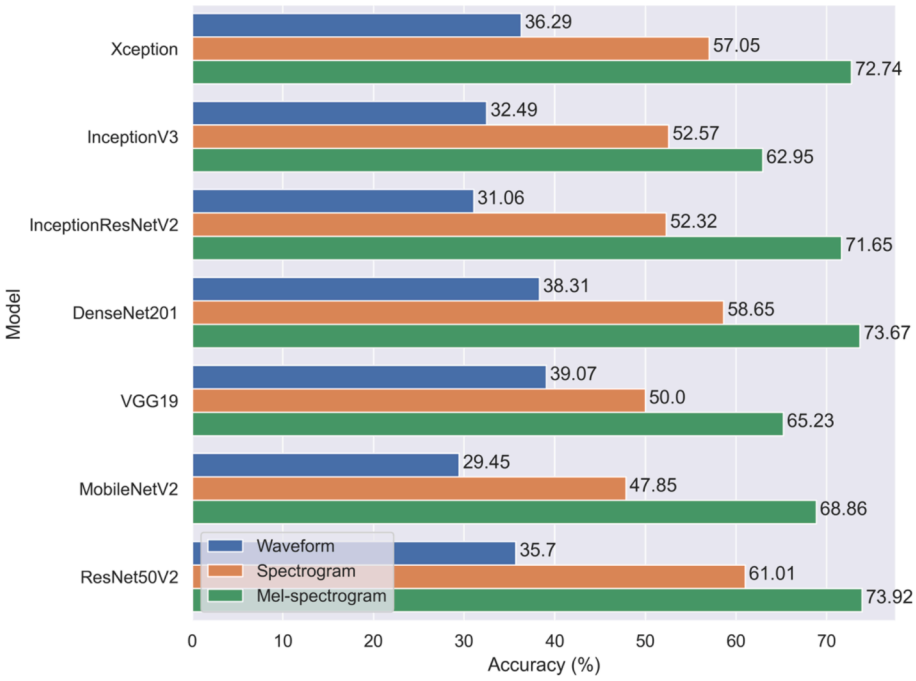


Fig. 6. The comparison of the sound visualization techniques, namely, (*i*) waveform, (*ii*) spectrogram, and (*iii*) mel-spectrogram.

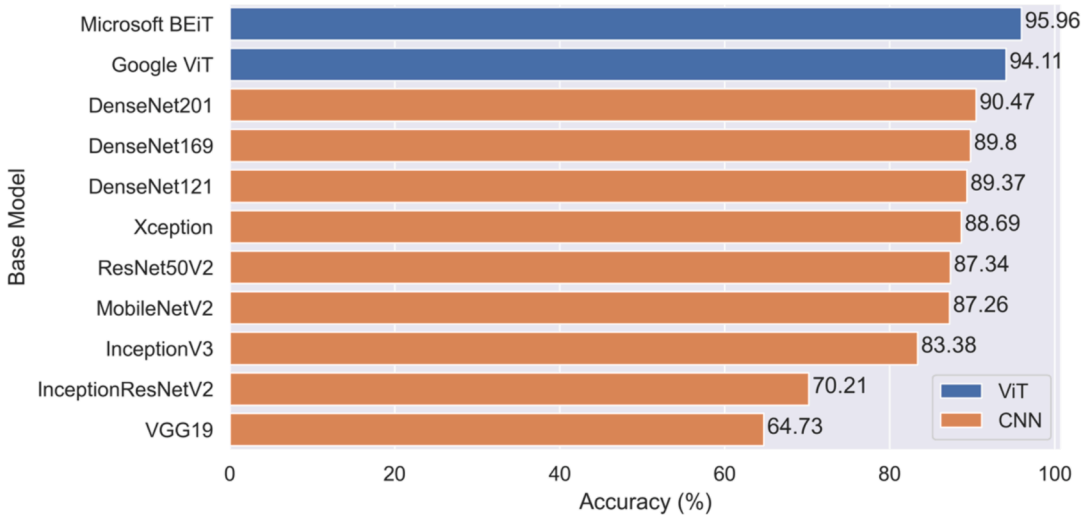


Fig. 7. The accuracy obtained by the proposed models on the test set.

Table 5
A comparison of the related studies that used the same dataset as the proposed study.

Related Work	Employed Technique(s)	Limitation(s)	Classification Accuracy
[16]	Traditional ML models	Lack of employment of more complex models and low accuracy	87.76%
[1]	FDAP as feature extractor and CDBN as classifier	Low accuracy	91.13%
[32]	Deep Neural Network	Low accuracy	88.47%
The proposed study	**ViT**	–	95.96%

work.

The inference time of the proposed classifier, the time it takes for a trained model to process a single input sample and produce an output, was calculated as low as 2.841 milliseconds, which indicates that the model can classify sounds very quickly and accurately without requiring excessive computational resources. This is particularly important in applications such as real-time wildlife monitoring, pet behavior analysis, and interactive systems involving animals.

5. Conclusion

Understanding animal sounds offers numerous benefits to humans in various fields, such as security, natural disaster prediction, ecological research, bioacoustics, precision agriculture, and search and rescue operations. Cats are among the most beloved pets worldwide. By comprehending cat sounds, we can unlock many advantages for those who

interact with them. Consequently, there is a need for a robust and fast cat sound classifier. In this study, we investigated the efficacy of both ViT and CNN architectures for classifying cat sounds represented as images. A gold standard dataset, named *CatSound*, was used to train and evaluate the proposed ViTs and CNNs. *CatSound* comprises 5,922 cat sound recordings categorized into 10 classes as follows: (*i*) warning, (*ii*) angry, (*iii*) defense, (*iv*) fighting, (*v*) happy, (*vi*) hunting mind, (*vii*) mating, (*viii*) mother call, (*ix*) paining, and (*x*) resting. The cat sound recordings were transformed into images using a sound visualization technique called mel-spectrogram, which was chosen based on the results of an experiment conducted specifically for this purpose. Our experimental results revealed that models based on ViT outperformed those based on CNN, achieving superior accuracy levels in categorizing diverse cat vocalizations. Specifically, the proposed model based on *Microsoft*'s *BEiT* outperformed the state-of-the-art by obtaining an outstanding accuracy of 96.95%. This underscores the efficacy of transformer-based approaches in capturing complex patterns within audio data, particularly when applied to the task of cat sound classification. Although CNN-based models also exhibited competitive performance, they were outperformed by ViT-based counterparts in terms of accuracy. This suggests that ViT architectures possess inherent advantages in extracting and representing relevant features from mel-spectrogram representations of cat sounds. The success of our proposed ViT models not only contributes to the advancement of animal sound classification techniques but also highlights the potential of transformer architectures in audio classification tasks. These findings underscore the importance of exploring and leveraging diverse neural network architectures to achieve optimal performance in decoding complex animal communication systems. It is worth mentioning that the calculated inference time of the proposed classifier stands at a mere 2.841 milliseconds, signaling its ability to swiftly and accurately classify sounds without imposing significant computational demands. This attribute holds particular significance in applications like real-time wildlife monitoring, pet behavior analysis, and interactive systems involving animals.

Future research endeavors could focus on further refining ViT-based models for cat sound classification, as well as investigating their generalizability to other animal species and environmental conditions. Additionally, exploring techniques to enhance the interpretability and robustness of these models would be valuable for real-world applications in wildlife monitoring and conservation efforts.

6. Ethics approval statement

This research article does not require ethics approval as it does not involve human participants, animal subjects, or sensitive personal data. The study solely utilizes publicly available datasets and existing literature for analysis and does not involve any experimental procedures or interventions.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Enver Kucukkulahli: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Abdullah Talha Kabakus:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study are publicly available upon a reasonable request to the data providers.

References

[1] Pandeya YR, Kim D, Lee J. Domestic cat sound classification using learned features from deep neural nets. Appl Sci (Switzerland) 2018;8:1–17. https://doi.org/10.3390/app8101949.
[2] Clutton-Brock J. Cats in ancient times. The British Museum Book of Cats Ancient and Modern. London: British Museum Press; 2002. p. 26–49.
[3] Bradshaw JWS. The cat-human relationship. The Behaviour of the Domestic Cat. CAB International; 1992. p. 163–76.
[4] Megna M. Pet Ownership Statistics 2024. accessed January 24, 2024 Forbes 2024. https://www.forbes.com/advisor/pet-insurance/pet-ownership-statistics/.
[5] cat - Explore - Google Trends. Google 2024. https://trends.google.com/trends/explore?date=2014-01-01%202024-01-25&q=cat&hl=en (accessed January 25, 2024).
[6] Wang X, Zhao Y, Pourpanah F. Recent advances in deep learning. Int J Mach Learn Cybern 2020;11:747–50. https://doi.org/10.1007/s13042-020-01096-5.
[7] Brown SL. The Social Behaviour of Neutered Domestic Cats (Felis catus). University of Southampton; 1993. Ph.D. Thesis.
[8] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016), Savannah, GA, USA: 2016, p. 265–83.
[9] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the Thirty-third Conference on Neural Information Processing Systems (NIPS 2019), Vancouver, BC, Canada: 2019, p. 8026–37.
[10] Wang AL-C. An industrial strength audio search algorithm. Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR 203), Baltimore, Maryland (USA), 26-30 October 2003 2003. https://doi.org/10.1109/IITAW.2009.110.
[11] Haitsma J, Kalker T. A highly robust audio fingerprinting system. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2002), Paris, France: 2002.
[12] Nanni L, Brahnam S, Lumini A, Barrier T. Ensemble of Local Phase Quantization Variants with Ternary Encoding. In: Local Binary Patterns: New Variants and Applications. Berlin, Germany: Springer; 2014. p. 177–88. https://doi.org/10.1007/978-3-642-39289-4_8.
[13] Salamon J, Bello JP, Farnsworth A, Kelling S. Fusing shallow and deep learning for bioacoustic bird species classification. Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA: IEEE; 2017, p. 141–5. https://doi.org/10.1109/ICASSP.2017.7952134.
[14] Manocha P, Badlani R, Kumar A, Shah A, Elizalde B, Raj B. Content-based representations of audio using siamese neural networks. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), vol. 2018- April, Calgary, AB, Canada: IEEE; 2018, p. 3136–40. https://doi.org/10.1109/ICASSP.2018.8461524.
[15] Nanni L, Rigo A, Lumini A, Brahnam S. Spectrogram classification using dissimilarity space. Appl Sci (Switzerland) 2020;10:1–17. https://doi.org/10.3390/APP10124176.
[16] Pandeya YR, Lee J. Domestic cat sound classification using transfer learning. International Journal of Fuzzy Logic and Intelligent Systems 2018;18:154–60. https://doi.org/10.5391/IJFIS.2018.18.2.154.
[17] Ntalampiras S, Ludovico LA, Presti G, Prato Previde E, Battini M, Cannas S, et al. Automatic classification of cat vocalizations emitted in different contexts. Animals 2019;9:1–14. https://doi.org/10.3390/ani9080543.
[18] Ntalampiras S, Kosmin D, Sanchez J. Acoustic classification of individual cat vocalizations in evolving environments. Proceedings of the 2021 44th International Conference on Telecommunications and Signal Processing (TSP 2021), Brno, Czech Republic: IEEE; 2021, p. 254–8. https://doi.org/10.1109/TSP52935.2021.9522660.
[19] Sun W, Lu V, Truong A, Bossolina H, Lu Y. Purrai: A deep neural network based approach to interpret domestic cat language. Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA 2021), Pasadena, California, USA: IEEE; 2021, p. 622–7. https://doi.org/10.1109/ICMLA52953.2021.00104.
[20] Nanni L, Maguolo G, Paci M. Data augmentation approaches for improving animal audio classification. Ecol Inform 2020;57. https://doi.org/10.1016/j.ecoinf.2020.101084.
[21] Yang Y, Sinnott RO. Automated recognition and classification of cat pain through deep learning. Proceedings of the International Conference on Big Data Intelligence

and Computing (DataCom 2022), vol. 13864 LNCS, Denarau Island, Fiji: Springer; 2022, p. 230–40. https://doi.org/10.1007/978-981-99-2233-8_17.
[22] Feighelstein M, Henze L, Meller S, Shimshoni I, Hermoni B, Berko M, et al. Explainable automated pain recognition in cats. Sci Rep 2023;13:1–16. https://doi.org/10.1038/s41598-023-35846-6.
[23] Ko K, Park S, Ko H. Convolutional feature vectors and support vector machine for animal sound classification. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2018- July, Honolulu, HI, USA: IEEE; 2018, p. 376–9. https://doi.org/10.1109/EMBC.2018.8512408.
[24] Sasmaz E, Tek FB. Animal sound classification using a convolutional neural network. Proceedings of the 3rd International Conference on Computer Science and Engineering (UBMK 2018), Sarajevo, Bosnia and Herzegovina: IEEE; 2018, p. 625–9. https://doi.org/10.1109/UBMK.2018.8566449.
[25] Weninger F, Schuller B. Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic: IEEE; 2011, p. 337–40. https://doi.org/10.1109/ICASSP.2011.5946409.
[26] Lu J, Ma R, Liu G, Qin Z. Deep convolutional neural network with transfer learning for environmental sound classification. Proceedings of the 2021 International Conference on Computer, Control and Robotics (ICCCR 2021), Shanghai, China: IEEE; 2021, p. 242–5. https://doi.org/10.1109/ICCCR49711.2021.9349393.
[27] Li Y, Wu Z. Animal sound recognition based on double feature of spectrogram in real environment. Proceedings of the 2015 International Conference on Wireless Communications and Signal Processing (WCSP 2015), Nanjing, China: IEEE; 2015, p. 1–5. https://doi.org/10.1109/WCSP.2015.7341003.
[28] Mohmmad S, Afroz M, Niharika P, Rahul RS, Rishitha L. Detection of wild animals using their sound for identification and conservation in the forest by implementing deep learning. Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES 2023), Chennai, India: IEEE; 2023, p. 1–7. https://doi.org/10.1109/ICSES60034.2023.10465272.
[29] Vithakshana LGC, Samankula WGDM. IoT based animal classification system using convolutional neural network. Proceedings of the 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE 2020), Colombo, Sri Lanka: IEEE; 2020, p. 90–5. https://doi.org/10.1109/SCSE49731.2020.9313018.
[30] Sharma S, Phoga A, Kadyan V. Automated classification of animal vocalization into estrus and non-estrus condition using AI techniques. Proceedings of the 21st International Conference on Information Technology, Proceedings (OCIT 2023), Raipur, India: IEEE; 2023, p. 553–8. https://doi.org/10.1109/OCIT59427.2023.10431114.
[31] Saleem N, Gao J, Irfan M, Verdu E, Fuente JP. E2E–V2SResNet: Deep residual convolutional neural networks for end-to-end video driven speech synthesis. Image Vis Comput 2022;119:1–10. https://doi.org/10.1016/j.imavis.2022.104389.
[32] Nanni L, Brahnam S, Lumini A, Maguolo G. Animal sound classification using dissimilarity spaces. Appl Sci (Switzerland) 2020;10:1–18. https://doi.org/10.3390/app10238578.
[33] McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. Proceedings of the 14th Python in Science Conference (SciPy 2015), Austin, Texas, USA: 2015, p. 18–25. https://doi.org/10.25080/majora-7b98e3ed-003.
[34] Chollet F. Keras: the Python deep learning API 2024. https://keras.io (accessed March 7, 2024).
[35] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (PMLR 9), vol. 9, Sardinia, Italy: 2010, p. 249–56.
[36] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Vienna, Austria: 2021.
[37] Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-training of image transformers. Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual: 2022.
[38] Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: a large-scale hierarchical image database. Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA: IEEE; 2009, p. 248–55. https://doi.org/10.1109/cvpr.2009.5206848.
[39] Hugging Face – The AI community building the future. Hugging Face 2024. https://huggingface.co (accessed March 4, 2024).
[40] Wolf T, Debut L, Sanh V, Chaumond J. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv 2019;1910(03771):1–8.
[41] Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: a novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 2018;18:6765–816.