International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Literature Survey on Bird Call Identification and Monitoring

Dr. Chethan B K[1], K R Raksha [2], Nikhitha S K [3], Shreya R [4], Soundarya G Rao [5]

[1]Associate Professor, CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.
[2]B.E Graduate (IV year), CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering Mysuru, Karnataka, India.
[3]B.E Graduate (IV year), CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering Mysuru, Karnataka, India.
[4]B.E Graduate (IV year), CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering Mysuru, Karnataka, India.
[5]B.E Graduate (IV year), CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering Mysuru, Karnataka, India.

Abstract - Birds play a main role in ecological monitoring, serving as indicators of environmental health. Many biological monitoring projects rely on the acoustic detection of birds. Despite the major increase in large datasets, this detection is often manual or semi-automatic, requiring manual tuning/postprocessing. The consistently used datasets for bird call monitoring and identification are [1] Cornell Bird Challenge(CBC) - 2020 dataset and [2] Xeno-Canto dataset. The major models used are [3] YamNet model which is a pre-trained model provided by the TensorFlow team. YamNet takes in a waveform of the given sound data sample and predicts the probability of each class, [1] ResNet-50, a deep CNN architecture for automated bird call recognition where spectrograms (visual features) extracted from the bird calls using Deep-CNN were used as input for ResNet-50 and Dmitry Konovalov et. al. in [2] suggested two approaches. The first approach was a stand-alone model trying the whole audio clip as input for ImageNet, ResNet, and VGGnet models. The second approach was a hybrid model that used window slides of raw audio, taking the spectrogram of each slide as input for CNN for representation and RNN for temporal correlation. The hybrid model achieved more accuracy. Even though many researchers have tried to automate bird call recognition, the desired accuracy of prediction is not achieved. On average the accuracy fluctuates between 60% - 72%. Additionally, prediction mainly depends on the quality of the dataset, the quantity of the dataset, the training pattern, and the input type given to the model.

***Key Words*:** Audio Classification, Birdcall monitoring, Bird Call identification, YamNet, ResNet-50, Deep-CNN, RNN, Spectrogram, Temporal Correlation.

1. INTRODUCTION

BirdLife International has expressed concern over the swift decline of global bird species. [8] According to a recent article by the organization, nearly half of all bird species have seen a significant population decrease, with only six percent experiencing an increase. Shockingly, one in eight species now faces the threat of extinction, and human activities are the primary cause. [8] Agriculture and industrialization, encroaching on crucial bird habitats, and using machinery and chemicals harmful to bird populations, are major contributors. Climate change is another significant threat affecting 34 percent of endangered bird species, with the potential for increased storms, wildfires, and droughts.

To identify and protect endangered bird species, one proposed solution is the analysis of their calls. Each bird has a unique vocalization that can possibly be used for identification. Traditional bird monitoring and identification methods, relying on manual observations and field surveys by ornithologists or birdwatchers, are time-consuming and labor-intensive. Monitoring large areas or tracking bird populations over extended periods is challenging. Automated audio identification and monitoring, employed in bioacoustics, use advanced technologies to recognize, analyze, and interpret sounds in diverse environments.

Audio classification, applying machine learning and signal processing techniques, categorizes audio data based on inherent characteristics. This enables systems to identify, organize, and respond to different sound patterns. Audio classification is a powerful tool in bird acoustics, transforming the study of avian ecosystems and biodiversity. Automated audio classification is increasingly vital for researchers and conservationists as traditional manual identification methods prove laborious and time-consuming. The technology's application can be further improved by enhancing prediction accuracy.[8]

2. METHODS

The research in [1] introduces a deep learning approach for predicting and analyzing bird acoustics using a dataset containing recordings of 100 bird species. The study explores various neural network architectures, including hybrid models that combine CNNs with other CNNs or RNNs such as LSTM, GRU, and LMU. Input processing involves a sliding window mechanism with a 500 ms window length and a 250 ms hop length. The temporal correlation block's output feeds into the final classification block, implementing a fully connected MLP with a layer of 512 neurons using ReLU non-linearity, followed by a dropout layer. The output layer matches the dataset's class count, providing conclusive classification outcomes. These findings, credited to the researcher, highlight the efficacy of the proposed hybrid models in bird sound classification, with performance evaluation focusing on accuracy and demonstrating the superiority of the hybrid model incorporating LMUs. The best-performing model achieves noteworthy accuracy across 100 bird species, showcasing the potential of the approach. [1]

Studies in [2] address a critical need in environmental monitoring by proposing an innovative approach for automated bird call recognition using ResNet-50, an architecture known for its success in computer vision tasks. The authors, Dmitry Konovalov et. al., leverage ResNet-50's proficiency, applying it to a publicly available dataset encompassing bird calls from 46 species. The meticulous methodology involves spectrogram extraction from bird calls as visual features, serving as ResNet-50 input. Results show promising accuracy levels, ranging from 60% to 72%, demonstrating the potential of deep learning in automating bird call recognition. These findings shed light on ResNet-50's capabilities and limitations in this domain, emphasizing the importance of considering call syllables as fundamental recognition units and suggesting future exploration into integrating audio and visual features for enhanced accuracy. [2]

The researchers, Ievgeniia Kuzminykh et. al., in [3] use trained models like YamNet, AlexNet and ResNet-50 to handle audio grouping jobs as well as tasks with time intervals. The models are trained to sort audio sounds into predefined classes. They use melgrams as their input representations. Two tests evaluate model efficacy: one centered around audio classification and the other involving interval retrieval based on a natural language query. The studies show that YamNet does better comparatively. The strengths being its ability to identify sounds and find the intervals within audio parts. The paper mostly focuses on using models pre-trained for audio classification and interval retrieval.[3]

In [4], Shwetank Choudhary et al. carefully treats the audio set. They reduce clips to 16 kHz and use one-second parts as input in it. Using the lightweight YAMNet to pick out features, a thick layer is thoughtfully added. It changes the output size. The Wave Encoder, a two-way network made of LSTM for raw waveform input, pairs features over time using two Bi-LSTM layers with 128 units each. The suggested shared plan connects Wave Encoder and already trained YAMNet features easily by putting them together. It uses a loss for training called binary cross-entropy. The study introduces two methods for adjusting time-related features—based on a measure of similarity and Bahdanau's method with learning weights. It tests the effect of these schemes carefully. Looking at the FSD50K data, the suggested system gets a mean average precision (mAP) of .445 on-device while using only 4.5MB memory. It significantly beats the earlier standard by 22%. [4]

In [5], Kaiming He et al. bring a new deep-learning method for picture recognition into play. They smartly change how layers work by using residual functions that use prior layer inputs to make things better. The study looks at how well residual networks (ResNets) work on ImageNet. It goes up to 152 layers deep and does a big study on the CIFAR-10 dataset to check if it scales well. The paper focuses on basic network structures like VGG nets, using 3x3 filters, direct downsampling with a stride of 2, and ending with global average pooling and a fully connected softmax layer. Furthermore, the study looks at using a square matrix Ws in the leftover function for sizes that match up and adding possible layers. These discoveries give important knowledge about the improvements in deep learning for image identification. [5]

In [6], the researchers, Palanisamy et. al., use machine learning from deep CNN models trained on ImageNet to create strong starting points for classifying sounds. The research uses a big chart to show how CNNs use spectrograms to learn. It also checks why the performance of these systems may change with random starting points and the order in which tiny batches are grouped. Models that are put together several times using the already trained DenseNet make a clear overall improvement in accuracy. Testing on ESC-50 and UrbanSound8K databases gives top-notch results. Important results point out the benefits of starting with weights that are already learned, finding elements causing different levels of performance, and showing how using more than one model in a team helps improve accuracy. [6]

2.1 Dataset

The dataset used in bird acoustics research plays a major role in advancing our understanding of avian vocalizations, behaviors, and biodiversity. These datasets are curated collections of audio recordings capturing the diverse sounds produced by various bird species. These recordings usually come from different environments and geographical places, giving a complete picture of bird song variety. Researchers and ornithologists rely on meticulously curated datasets to train and evaluate machine learning models for automated bird sound identification. These datasets typically include labeled audio samples, where each recording is associated with information about the bird species producing the sound. The variety within the dataset ensures that the models generalize well across different species and environmental conditions. It's important to select a good set of data because it makes machine learning models more accurate and useful in bird sound study, helping ornithologists discover new things and save birds' habitats.

Gaurav Gupta and colleagues [1] utilized the Cornell Bird Challenge (CBC) 2020 dataset, comprising 264 bird species, each accompanied by varying audio samples ranging from 9 to 1778 for every species. This dataset contains recordings capturing multiple and potentially overlapping bird vocalizations amidst background noise, including up to 100 different bird species. Due to limited samples in some classes, 100 bird classes were selected based on having the highest number of samples. To ensure balance within each class, efforts were made to guarantee that each class contained a minimum of 100 samples.

Dmitry Konovalov et. al. in [2] have used a publicly available dataset that is a subset of Xeno-Canto, developed by Nanni et al. 2016, from the Xeno-Canto website, which consists bird call audio within a radius of up to 250 km of the city of Curitiba, in the South of Brazil. They removed all bird species with less than 10 audio samples. After these filters, 2814 audio

samples, representing 46 bird species remained in the dataset with the sample rate of audio being 22.05 KHz.

Ievgeniia Kuzminykh et. al. in [3] have used a database that consists of an expanding ontology of 527 audio event classes and an assemblage of 2,084,320 human-labelled 10-second sound clips taken from YouTube videos. Both training and evaluation datasets were merged and then randomly split 80/20 for the training and testing subsets. Each of the 6 classes has approximately 120 audio samples linked to them in the training data that had been picked and then another 11266 augmentations had been generated so that each class is represented by approximately 2000 samples in the training dataset. 12000 were used for training, of which 734 were 10-second raw samples in WAV format.

Palanisamy et. al. in [6] have used ESC-50, UrbanSound8K, and the GTZAN dataset.

The ESC-50 dataset comprises 2000 audio clips with a duration of 5 seconds each, belonging to 50 different classes of environmental sounds. The audio clips are uniformly sampled at a rate of 44.1kHz. The dataset, divided into five parts, and the accuracy of the classification is determined by evaluating the model's performance through cross-validation on all the divisions. The sounds included in the ESC-50 dataset range from chirping birds to car horn sounds.[6]

UrbanSound8k: The UrbanSound8k dataset consists of 8732 clips belonging to 10 classes of different urban sounds. Each audio clip is of length <= 4s, and the sampling rate varies from 16kHz to 44.1kHz. We resampled all audio clips to a sampling rate of 22.5kHz. The dataset is officially split into 10 folds, and cross-validation is performed on these 10 folds.

GTZAN Dataset: The GTZAN dataset consists of 1000 music clips each of length 30 seconds. There are 10 distinct genre classes. The music clips are sampled at a rate of 22.5kHz. There is no official training and validation split of the dataset therefore we used 20% of the original data for validation with an equal number of samples for each class and the rest of the data for training.

2.2 Working

The hybrid mode of audio signal processing described in [1] utilizes the sliding window mechanism, which is specifically designed for bird sound classification. This mechanism traverses raw audio clips with a sliding window of a specified length (Ws) and hop length (Hs), thereby capturing temporal components and generating 26 slides from a 7-second audio clip. Each slide results in a 128x32 single-channel 2-dimensional mel-spectrogram input, which is processed using three integral blocks - Representation, Temporal Correlation, and Classification. The Representation block employs a CNN to extract representative features from input slides, forming a two-dimensional array by concatenating these features, which is then used as input for the Temporal Correlation block. The resulting output from the block is directed to the final classification block, which produces softmax outputs. The schematic representation of this hybrid methodology is illustrated in Figure 1 [1], providing a detailed visualization of the model's architecture. This hybrid approach is highly effective in analyzing the temporal aspects of audio signals and has been praised by researchers in [1] for its efficiency in bird sound classification.

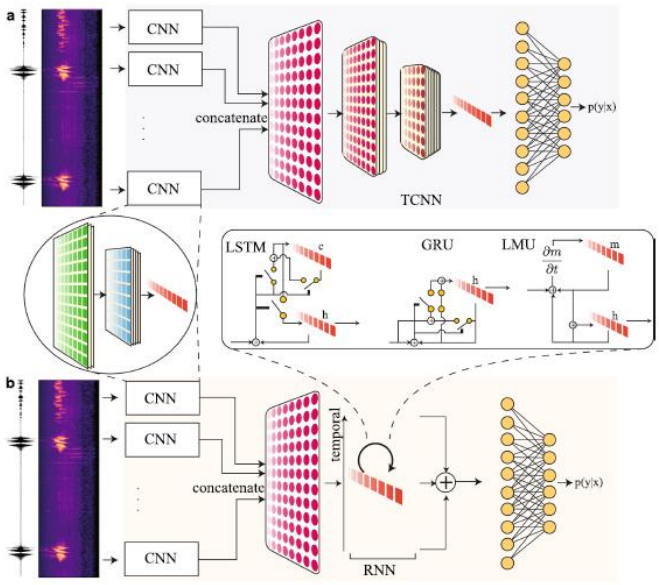

Fig -1: *"An outline of mixed models for categorization. The model involves using a CNN to represent the data and either a TCNN in (a) or an RNN in (b) for temporal correlation extraction. The CNN outputs are combined before being supplied to the temporal layer in both cases, (a) and (b)." [1]*

According to [2], residue learning is used for audio classification. The Inception-V4 architecture is known to utilize this residue learning. CNNs face the issue of decreasing accuracy when larger layers are used. However, this drawback can be overcome with the use of residual learning. The method of residual learning involves shortcut connections that directly connect the input to some other subsequent layers for training the CNN. In the case of Bird Call Recognition, ResNet's high performance for image classification made it a suitable candidate for audio classification too. To take advantage of ResNet-50's Residual Learning capability, researchers chose to use it for the task. [2] ResNet-50's use of Residual Learning, the researchers went forth with ResNet-50. A representation of a unit of residual learning is depicted in Figure 2. [2]

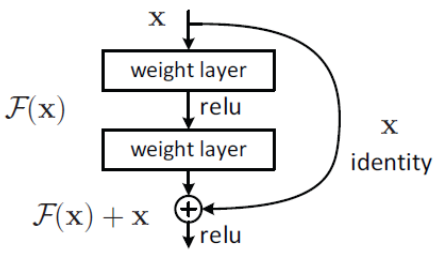

Fig -2: *Unit of Residual learning [8]*

The study described in [3] involves the application of YamNet, AlexNet, and ResNet-50 for audio event classification. Based on the AudioSet-YouTube corpus [3], YamNet predicts 521 audio event classes. However, six classes out of 527 were excluded due to recommendations from the reviewers. The output of the model usually consists of 521 classes. For the experiment's purposes, the array of 521 elements was replaced with a 6-element array, where each element represents the probability of one of the 6 classes selected for the study. The output was then passed through a Random Forest classifier, resulting in the final output used for

interval retrieval, as the model performed better with the classifier than on its own. [3]

Similarly, AlexNet and ResNet-50 were both modified to output an array of 6 classes instead of the default 1000 classes. AlexNet originally had 8 layers, and ResNet-50 had 50 layers. Following their modification, both models were fitted with a Random Forest classifier on top of their outputs.[3]

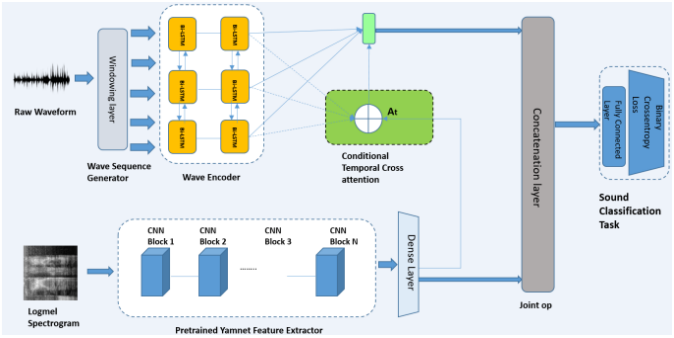

Fig -3: *The proposed LEAN model [4]*

In [4], the researcher examined various large-scale audio dataset-based pre-trained models for transfer learning, ultimately selecting YAMNet over Vggish and PANN due to its efficiency in memory and computation.[4] YAMNet is known for its lightweight nature, boasting nearly 3.7 million weights and achieving .306 mAP on the evaluation set after being trained on the AudioSet dataset. Unlike its competitors, YAMNet is well-suited for resource-constrained environments such as mobile phones and edge devices. To further enhance overall model efficiency, a dense layer of 256 units (the projection layer) was added after the 1024 embedding vector of YAMNet, as shown in Figure 3. [4] This modification fine-tunes the Wave Encoder output and reduces the joint embedding size. Finally, to calculate the affinity scores of time sequences from Wave Encoder, a simple dot product is performed with the YAMNet output. [4]

After conducting an in-depth analysis, Kaiming He et al concluded that several methodologies can be employed to improve the performance of deep neural networks, particularly in image classification using the ImageNet dataset. [5] They examined both Plain Networks and Residual Networks (ResNets) and found that the deeper network exhibited higher training errors despite employing Batch Normalization, commonly known as "degradation." To address this issue, the authors introduced Residual Networks, which was previously hinted at, which include shortcut connections that facilitate the flow of information. The study compared the 18-layer and 34-layer occasional max-pooling layers to reduce grid resolution. Plain Networks with the 34-layer ResNet, highlighting the superior performance of the latter and demonstrating the efficacy of residual learning in mitigating the degradation problem. The ImageNet architecture is depicted in Figure 4.

Regarding [6], Palanisamy, K et al utilized three established models, which were trained on ImageNet, to address their research challenge. [6] The first model used was Inception, with a unique architecture featuring an Inception layer. This layer is composed of 1x1, 3x3, and 5x5 convolutional layers that are combined to form a single vector. Multiple Inception layers are then stacked together, with occasional max-pooling layers to reduce grid resolution. The second model employed was ResNet, which is characterized by a series of residual blocks. Each block consists of two 3x3 convolutional layers, batch normalization, and ReLU activation. The input is then added to the output of the ReLU activation function using a skip connection for identity mapping. Finally, the third model used was DenseNet, depicted in Figure 5, which uses a dense convolutional network structure that connects each layer to every other layer in a feed-forward manner. Unlike traditional networks, DenseNet builds L(L+1)/2 direct connections, with each layer using the preceding layer's feature maps as inputs and sharing its feature maps with subsequent layers. The authors' strategic approach to utilizing diverse architectures, such as Inception, ResNet, and DenseNet models, in addressing their research challenge is evident in their study. [6]

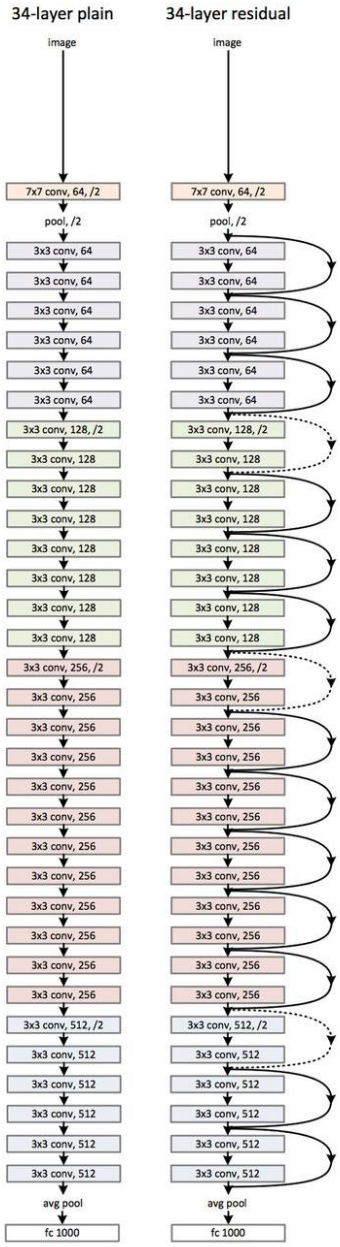

Fig -4: *"Example network architectures for ImageNet. Left: a plain network with 34 parameter layers. Right: a residual network with 34 parameter layers. The dotted shortcuts increase dimensions" [5]*

2.3 Characteristics of the methods

CNNs or Convolutional Neural Networks [1] have played a crucial role in audio classification tasks. They have unique attributes that enable them to work efficiently with sound data, especially in spectral analysis. Their hierarchical feature learning abilities make it possible to automatically identify patterns in the frequency domain. As a result, CNNs are ideal for tasks such as audio event recognition and sound classification.

Recurrent Neural Networks (RNNs) [1] are a type of neural network that can be used to classify audio data with sequential dependencies. Popular variants of RNNs like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are particularly useful for capturing the temporal context of audio sequences, making them ideal for speech recognition, music genre classification, and other audio processing applications. One of the key advantages of LSTMs and GRUs is their ability to handle long-term dependencies, which ensures that information is preserved over extended audio sequences, making them even more useful in a variety of audio processing scenarios.

Liquid State Machines (LMU) [1] is a unique approach to dynamic audio processing that draws from liquid computing principles. These machines are particularly effective at capturing the temporal dynamics present in audio signals, which is a valuable asset when dealing with features that evolve. LMUs prioritize memory and context, making them ideal for applications that require a nuanced understanding of recurring audio motifs and patterns. In the realm of audio classification, LMUs can enhance the quality of models by accounting for the complex dynamics present in sound data.

ResNet-50 [2] is a notable model for audio classification due to its deep and efficient architecture that utilizes skip connections to overcome the vanishing gradient problem. With this design, the model can effectively learn hierarchical audio

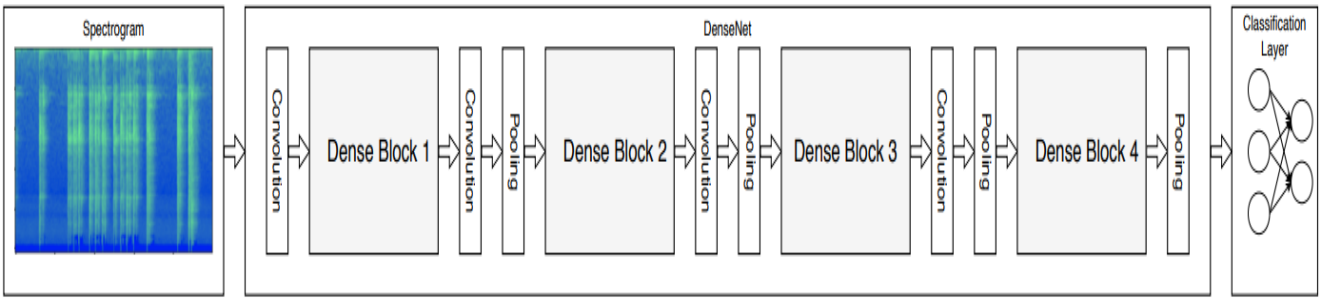

Fig -5 : DenseNet architecture *[6]: In each Dense Block, several convolution layers are used, and their inputs include features from all the previous layers within that block. The chosen architecture is DenseNet 201, which organizes these blocks with {6, 12, 48, 32} convolution layers in sequence.*

features, making it successful in various classification tasks. Additionally, ResNet-50 has strong transfer learning capabilities, allowing it to apply knowledge gained from image datasets to smaller audio datasets. Its scalable performance and versatility make it a powerful solution for different audio classification challenges, highlighting its importance in advancing the state-of-the-art in this field.

YamNet [2] is an advanced deep-learning model created by Google that is designed for audio event classification. It utilizes a deep convolutional neural network to accurately analyze the spectral features of audio signals. With its extensive training on a diverse dataset, it can identify and categorize a broad range of audio categories. As a result, it is a dependable choice for real-world audio classification applications.

AlexNet [3], initially developed for visual recognition tasks, has demonstrated its versatility and feature extraction abilities in the field of audio classification. Although it is not as specialized as YamNet in audio tasks, AlexNet's architecture, which includes both convolutional and fully connected layers, enables it to learn distinctive features from audio spectrograms automatically.

The Wave Encoder [4] is built to handle raw waveform input and uses a two-way network with Long Short-Term Memory (LSTM) units [1]. The model has two Bidirectional LSTM layers, each with 128 units, and is highly effective in capturing temporal dependencies in audio data. It pairs features over time to learn complex patterns in the raw waveform input. The LSTM layers work in both directions, making the model adaptable and skilled at capturing subtle temporal relationships. The Wave Encoder's use of LSTM units highlights its effectiveness for various audio processing tasks.

The DenseNet model [6], which is known for its densely connected blocks, has been proven to be effective in promoting direct connections between layers, resulting in efficient feature reuse. This design is particularly beneficial in audio

classification as it enables the model to extract and combine relevant features, therefore capturing complex patterns and dependencies in audio data. The Inception model [6] has proven to be highly effective for audio classification tasks due to its ability to extract features of varying scales from audio signals. This is achieved by using parallel convolutional layers with diverse kernel sizes, which enables the model to capture the temporal hierarchy of audio features. Furthermore, the Inception model is versatile in handling complex audio data and is capable of efficiently learning diverse features, which contributes to its effectiveness in audio classification applications.

2.4 Comparison of Studies Table 1 compares all the mentioned models in their respective papers and lists their strengths and weaknesses.

TABLE - 1: Comparative overview of Audio Classification Studies

Paper	Introduction	Characteristics of methods	Strengths	Weaknesses
Audio Interval Retrieval using Convolutional Neural Networks	Studies sound retrieval from natural language queries using pre-trained models, evaluating their accuracy in classifying audio and retrieving relevant intervals.	**Pre-trained Models:** YamNet (521 audio events), AlexNet (images), ResNet-50 (images) **Data Selection & Augmentation:** Carefully selects data from AudioSet (16902 samples) focusing on relevant classes. **Focus:** Emphasizes interval retrieval based on natural language queries. **Interval retrieval performance:** All models	Use of pre-trained models, YamNet architecture, focus on interval retrieval, natural language query support, Random Forest classifier augmentation.	Limited training dataset size, potential misclassifications, dependency on training data quality, and complexity of audio retrieval task.
LEAN: Light and Efficient Audio Classification Network	Explores efficient audio classification for resource-constrained devices. Combines transfer learning with YAMNet (pre-trained model) and a trainable Wave Encoder.	**Transfer Learning with Pretrained Models:** Leverages knowledge from YAMNet, pre-trained on AudioSet (large-scale dataset). **YAMNet Architecture:** MobileNet-based, efficient, and effective for audio classification. **Wave Encoder:** Trainable component capturing task-specific features from raw waveforms. **Combining Trainable and Pretrained Components:** The hybrid approach balances accuracy and resource efficiency.	Transfer learning with YAMNet boosts accuracy, and is strong with limited data. Wave Encoder enables task-specific adaptability. Resource-efficient and lightweight for deployment on constrained devices. Balances accuracy and efficiency.	Sensitive to highly divergent target domains where YAMNet features might not be suitable.Performance depends on the quality and representativeness of the pre-trained YAMNet. Transfer learning might be less effective for rare or poorly represented classes in YAMNet.
Deep Residual Learning: A Solution to Degradation in Deep CNNs for Image Classification	Addresses the limitations of deep CNNs: Vanishing/exploding gradients and degradation (accuracy declines as depth increases). Proposes deep residual learning as a solution.	**Residual Learning:** Learns differences between desired output and layer outputs. Shortcut connections bypass layers for better information flow and deeper networks. Bottleneck architectures further reduce complexity. **Network Architecture:** Built from repeating residual blocks with shortcuts. Stacking blocks creates deeper, more accurate networks. **Implementation:** Standard image pre-processing, batch normalization, etc. SGD training with momentum. 10-crop and multi-scale testing for evaluation.	Enables training of significantly deeper networks without accuracy loss. Eases optimization and leads to faster convergence. Mitigates degradation problem. Identity mappings provide efficiency. Bottleneck architectures further reduce complexity.	Projection shortcuts introduce extra parameters compared to identity mappings. Deeper networks require more resources.
Rethinking CNN Models for Audio Classification	Focuses on transfer learning with ImageNet-pre-trained CNNs for audio classification, exploring its effectiveness and	**Transfer Learning**: Leverages large pre-trained models (ResNet, DenseNet, Inception) on ImageNet. Uses mel-spectrograms as input. Single model & feature set achieve state-of-the-art performance.	State-of-the-art accuracy with simple model & features. Validates image-to-audio transfer learning.	Sensitive to hyperparameters. Limited understanding of transfer learning mechanisms.

	advantages.	**CNN Models:** Popular models like ResNet, DenseNet, and Inception are used with mel-spectrogram input. Deep ensembles of identically trained models improve performance.	CNN learning insights through Integrated Gradients.	Potential overfitting on small datasets.
Bird Call Recognition using Deep Convolutional Neural Network, ResNet-50	This paper explores deep learning, specifically ResNet-50 trained on images, to automatically recognize bird calls in field recordings with high accuracy. Enabling large-scale analysis of ecological data will help enhance the process of environmental monitoring.	**Model:** ResNet-50, a deep CNN trained on ImageNet, was chosen for its ability to handle large datasets and complex features. **Data**: 2814 audio samples from 46 bird species were extracted from Xeno-Canto and converted into spectrogram images. **Preprocessing:** Minimal preprocessing (dividing pixel intensity, subtracting mean) and image augmentation (random scaling) were used. **Model Configuration:** ResNet-50 was adapted for 46-class bird call classification with pre-trained ImageNet weights and a sigmoid activation function. **Training:** 100 hours on an Nvidia GTX 1070 GPU, monitoring accuracy and spectrogram length impact.	ResNet-50 deep learning model: Adapted from computer vision, achieving 60%-72% accuracy in recognizing calls from 46 bird species. Publicly available dataset: Enhances transparency and research collaboration. Spectrogram input: Captures visual features of bird calls for effective model use. Scalability and remote deployment: Suitable for handling large datasets and monitoring inaccessible areas.	**Pre-trained model bias:** Reliance on ImageNet-trained ResNet-50 raises concerns about adaptability and potential bias. **Sensitivity to call length:** Accuracy improvement with shorter spectrograms necessitates further investigation for varying call durations. **Limited species scope**: Generalizability to a broader range of bird species needs more consideration.
Comparing Recurrent Convolutional Neural Networks for large-scale Bird Species Classification	This paper delves into a hybrid deep learning model for automated bird call recognition, addressing the limitations of image-based classification. By combining convolutional and recurrent neural networks, the model captures both spatial and temporal features in bioacoustics data	**Input Representation:** Mel-spectrograms are chosen for audio signal representation due to their effectiveness. **Model Training:** Adam optimizer is used during training. 50 epochs of training are performed, selecting the best accuracy model for testing. **Stand-alone Models:** Utilize ImageNet-trained CNNs (VGG16, ResNet) as classifiers. Have an output layer with 100 neurons for 100 classes. **Hybrid Models:** Transform raw audio into mel-spectrograms. **Contain two main components:** Representation: CNNs extract features from spectrogram slides. Temporal Correlation: Options include CNNs or RNNs (LSTM, GRU, LMU). **Classification**: MLP with 512 neurons, dropout, and output layer.	**Transparency, and reproducibility**: Detailed methodology descriptions and publicly available code on GitHub. **Mel-spectrogram effectiveness**: The choice of mel-spectrograms for frequency transformation aligns well for bioacoustic data. **Hybrid model innovation:** The novel hybrid CNN-RNN architecture effectively captures both spatial and temporal dependencies in bird calls, addressing a common limitation of image-based models. **Comprehensive model exploration:** Experimentation with various CNN and RNN architectures in different model components ensures thorough performance evaluation. **LMU memory insights:** The	**Evaluation metrics**: Relying solely on accuracy overlooks valuable measures like precision and recall, especially for imbalanced data. **Hyperparameter tuning**: Lack of details on how hyperparameters were chosen and their impact on performance weakens the study. **Comparative analysis**: Missing comparison with existing state-of-the-art models in bioacoustics hinders understanding of the approach's significance. **Interpretability of results**: Clustering information alone without deeper analysis limits understanding of why certain species are grouped.

			analysis of the LMU mechanism offers a valuable understanding of its temporal behavior.	

3. DISCUSSION AND CONCLUSION

In the discipline of automatic detection of bird sounds, researchers have taken advantage of deep-learning techniques to address the daunting problems in sound recognition. One of the hybrids, as claimed in the reference [1], is a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as LSTM, GRU, and RMUs, to perpetrate the strong points of both spectral and temporal aspects of bird calls. Incorporation of LMU is paramount to detail the temporal attributes which are critical in understanding the dynamics. The other approach advocated in [2], employing ResNet-50, is known for the efficiency in residual learning; and is used to address the problem of the vanishing gradient. The model had been developed based on the transfer learning technique which means that it used the pre-existed knowledge from image datasets to recognize complicated bird call patterns.

For the application of pre-trained models [3] looks at the use of YamNet, AlexNet and ResNet-50 for audio event classification. YamNet is unique in its audio tasks specialization which helps it come up with more efficient results yield in terms of sound and interval classifications.

The [4] model that is used by the LEAN (Lightweight Echocardiogram Assessment Network) takes a time efficient approach by using YamNet having a lightweight design and working with bidirectional LSTM layers within a wave encoder where temporal dependencies are captured. Moreover, there is [6] that presents different kinds of pre-trained models (DenseNet, Inception, ResNet) for sound classification along with the focus on the need for different feature extraction methods. These researches have distinguished the importance of custom model architectures in the continuous improvement and refinement of current deep learning systems that work for accurate bird sound recognition.

REFERENCES

[1] Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S,. Ferres, J L. “Comparing recurrent convolutional neural networks for large scale bird species classification.” Sci Rep 11, 17085 (2021).

[2] Sankupellay, Mangalam & Konovalov, Dmitry. (2018). Bird Call Recognition using Deep Convolutional Neural Network, ResNet-50. 10.13140/RG.2.2.31865.31847.

[3] Kuzminykh, I., Shevchuk, D., Shiaeles, S., & Ghita, B. (2020). “Audio interval retrieval using convolutional neural networks.” In Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 20th International Conference, NEW2AN 2020, and 13th Conference, ruSMART 2020, St. Petersburg, Russia, August 26–28, 2020, Proceedings, Part I 20 (pp. 229-240). Springer International Publishing.

[4] Choudhary, S., Karthik, C. R., Lakshmi, P. S., & Kumar, S. (2022, November). “LEAN: Light and Efficient Audio Classification Network.” In 2022 IEEE 19th India Council International Conference (INDICON) (pp. 1-6). IEEE.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). “Deep Residual Learning for Image Recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Palanisamy, K., Singhania, D., & Yao, A. (2020). “Rethinking CNN Models for Audio Classification.” arXiv preprint arXiv:2007.11154.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[8] BirdLife. 2022. State of the World's Birds report

[9] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in Proceedings of the 23rd Annual ACM Conference on Multimedia. ACM Press, 2015, pp. 1015–1018.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708

[11] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in 22nd ACM International Conference on Multimedia (ACM-MM’14), Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Advances in neural information processing systems, 2014, pp. 3320–3328.

[13] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Advances in neural information processing systems, 2019, pp. 3347–3357

[14] Fonseca, Eduardo & Favory, Xavier & Pons, Jordi & Font, Frederic & Serra, Xavier. (2020). FSD50K: an Open Dataset of Human-Labeled Sound Events.

[15] https://research.google.com/audioset/ontology/index.html

[16] https://github.com/SarthakYadav/fsd50k-pytorch

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge.