“A Comprehensive Review Paper on Bird Sound Prediction Models and Techniques”

Neha Bhalerao
Department of Computer Engineering,
Government College of Engineering ,
Yavatmal , Maharashtra, India
bhaleraoneha56@gmail.com

Renuka Dhule
Department of Computer Engineering,
Government College of Engineering ,
Yavatmal , Maharashtra, India
renukadhule2002@gmail.com

Pallavi Kasture
Department of Computer Engineering,
Government College of Engineering ,
Yavatmal, Maharashtra, India
pallaviikasture@gmail.com

Saurabh Chaunde
Department of Computer Engineering,
Government College of Engineering ,
Yavatmal , Maharashtra, India
chaundes@gmail.com

Prof. Sudesh A. Bachwani
Department of Computer Engineering
Government College of Engineering Yavatmal, Maharashtra, India, Assistant Professor,
Dr. Babasaheb Ambedkar Technological University
Lonere, India.

Abstract : In this paper, we present a system that can recognize various bird species based on their sounds. It is a challenging feat to identify these animals using their sounds without the need for physical contact. We use a two-step procedure to accomplish this .To create a good dataset, we collected various bird species sounds and made improvements to the clips using various techniques. These included separating the parts into sections, taking silent parts out, and putting the pieces back together. This resulted in spectrograms, which are images for each sound clip.In the second step, we used a special kind of computer program called a neural network. We fed the spectrograms into this program as input. The program, which works sort of like a brain, used the pictures to figure out what kind of bird sound it was hearing. We also made a model that can do this in real-time, meaning it can work on sounds as they happen.

Keywords: **Bird species identification, bird sound, sound pre-processing techniques, ConvolutionalNeural Network, Spectrograms, Acoustic Sensors.**

INTRODUCTION

Deep learning is a type of machine learning that uses neural networks and representations. It can be supervised, semi-supervised, or unsupervised. The term "deep" refers to the multiple layers that are used in the network.

Various types of deep learning architectures are used in various fields, such as drug design, natural language processing, and speech recognition. They have been able to perform better than human experts in some cases.

An artificial neural network is modeled after the distributed communication networks in biological systems. It has various notable distinctions from living organisms' brains, which are characterized by analog and dynamic elements.[1]

BIRD SOUNDS

Birds and especially sounds produced by birds have always fascinated people. Many people enjoy watching birds. Also, knowing which bird is making a sound is useful for studying nature and the environment. For example, scientists use bird sounds to see how human actions affect animals. They listen to and count birds to understand their impact on the environment.

In the last decade many researchers have devoted considerable efforts towards automatic recognition of bird species. Recognition of bird species that produce tonal, harmonic, and inharmonic sounds were studied in A statistical manifold approach using supervised learning was raised to automatically identify which species of bird is present in an audio recording. In the publications cited here the feature extraction are mainly based on time domain or frequency domain. In light of this, in this paper we consider the spectrograms (time-frequency representation) as a feature parameter. Unlike music or sounds from the environment, bird songs are a type of natural sound that often sounds like a organized series of short sounds with specific meanings. These short sounds are usually called elements or syllables. What's interesting is that many bird songs have a collection of these syllables, each with its own characteristics. This is why bird songs can be so melodious. When we look at pictures of bird songs' sounds, called spectrograms, we can see that the arrangement of these syllables

makes the pictures look different for each bird species. This is why we need to focus on finding specific features in small parts of the sound instead of looking at the whole sound together. Thinking about this, we suggest a clever way to break down bird songs into smaller parts based on time and sound frequency. We split continuous bird songs into individual syllables. Each of these syllables is like a building block for recognizing bird sounds. We then look closely at each syllable and take out special details from them.[2]

.

BIRD SOUND DATASETS

In the context of exploring prediction methods and technology for bird sound recognition, data acquisition plays a fundamental role in capturing the intricate vocalizations of avian species. The process of data acquisition involves the collection of audio recordings in natural environments, which are then used as the basis for developing recognition algorithms and conducting various.

Data acquisition methods have evolved significantly due to advancements in sensor technology and recording equipment. In the past, researchers often relied on manual field observations and analog audio recording devices.

However, with the advent of digital audio recorders and automated sensor networks, data collection has become more efficient and capable of covering larger geographic areas.[3]

However, data acquisition in this field is not without challenges. Ethical considerations must be taken into account to avoid disturbing bird populations with the presence of recording devices. Additionally, data management and storage are critical aspects, as the amount of audio data collected can be substantial. This requires effective organization, storage, and archiving strategies to ensure the accessibility and usability of the recorded data.

In summary, data acquisition is the foundation upon which the exploration of bird sound recognition methods and technologies is built. As technology continues to advance, data acquisition techniques will play an integral role in providing the essential raw material for training machine learning models, refining recognition algorithms, and advancing our understanding of avian ecosystems and their dynamics.

Now, our dataset has 500 sound clips, which must be divided into a Train dataset, Validation dataset and Test dataset before being given as input to the CNN in the ratio of 70:10:20. The Train dataset is used to train the network and fit the model. Validation dataset is used to tune the hyperparameters of a model during iterative training. Test dataset is used to provide an equitable evaluation of the terminal model fit on the training dataset. Finally, the dataset can be divided into several segments and cross validation can be used to ensure that the sound clips present in each dataset have equal data representation and distribution from all classes.[4]

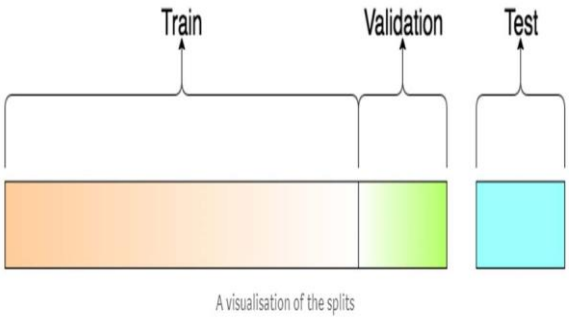

FIG 1.1: DIVISION OF THE DATASETS

LITERATURE REVIEW

Bird sound recognition has gained significant attention in recent years due to its applications in various fields, including ornithology, wildlife conservation, and biodiversity monitoring. This literature review provides an overview of the key developments, methods, challenges, and future prospects in the domain of bird sound recognition.[5]

METHODOLOGY

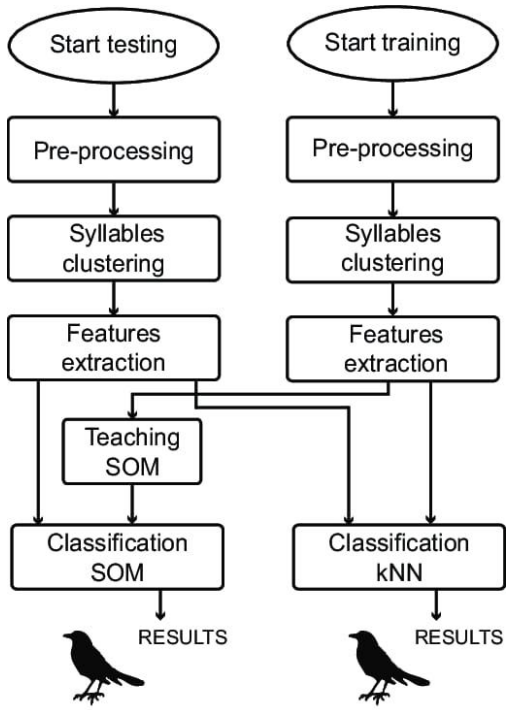

FIG 1.2: Flowchart of the Methodology

SELF ORGANIZING MAP

Self Organizing Map is type of Artificial Neural Network (ANN). This type of ANN learns without a teacher, using only the observation of the input data (unsupervised learning). Network map, which creates a static grid cell, has a fixed size. It usually has a rectangular or hexagonal structure. Weights of input neurons can be initiated with random values. SOM has two basic methods of changing the neurons weights. The

first one - Winner Takes All (WTA): the neuron, whose weights are closest to the input vector components is modified in such a way that its weights are as close as possible to the Input vector. The second one. Winner Takes Most (WTM): neuron with weight most similar to the input value is called the winner.[6]

KNN (K-NEAREST NEIGHBOUR ALGORITHM)

In KNN algorithm the recognition process involves calculating distances in parameters space X between the unknown , object and all objects from the training set ECU for k = 1, 2,..., I. where I is the number of training examples. In presented project Euclidean distance has been used:

$$\mathbf{d(x_j, x_k)} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}\mathbf{ij} - \mathbf{xki})^2}$$

Obtained distances are sorted in an ascending order. Object r, is assigned to this class, which is the most common among k nearest objects.[7]

CLASSIFIERS

Classification is an algorithm, which assigns objects to groups (called classes) based on object features. Features are usually presented in a vector. All features values in a task are called the training set CU.

PREPROCESSING

The goal of preprocessing is adaptation and simplification of the signal for further analysis. It is divided into three steps filtration, normalization and wavelet decomposition. The aim of filtration, done by the use of band-pass filter, was to remove higher frequencies.

After filtration data were normalized. The goal of normalization was to eliminate the influence of the amplitude from the further analysis. Different amplitudes may be the result of various conditions during signal registration. In this study signal was normalized to fit [-1,1] value interval. Unfortunately, normalization also decreased distances between classes. However, this was a necessary step, before proceeding to the next stages. After normalization wavelet analysis was used for signal de-noising. Noise usually comes from recording apparatus, as well as from the environment.[8]

DIVISION INTO SYLLABLES

The definition of syllable is a problem in phonetics and phonology of human speech. This problem becomes even greater when it comes to birds. Therefore, one of the biggest challenges of this study was syllable extraction. Physically the syllable is defined as a segment which has higher intensity than its neighborhood. In this paper, following considerations are bused on the signal time domain.

Division into syllables was divided into three parts. The first part was approximation, which reduced the noise and dimensionality of signal samples. After that local maxima and minima were designated, based on the gradient of signals polynomial approximation.

The syllables were clustered between two neighboring minima and usually had one maximum. However, if a time period of a syllable was to small or differences between extrema were to low (what means, that this observation is a part of the same syllable), it was added to the previous syllable. Values of factors in this algorithm have great influence on classifiers performance. At the beginning the values of factors were established basing on the observation of the system, and after that, factors were optimized basing on the highest results. All the research and the analysis was carried out on isolated syllables.[9]

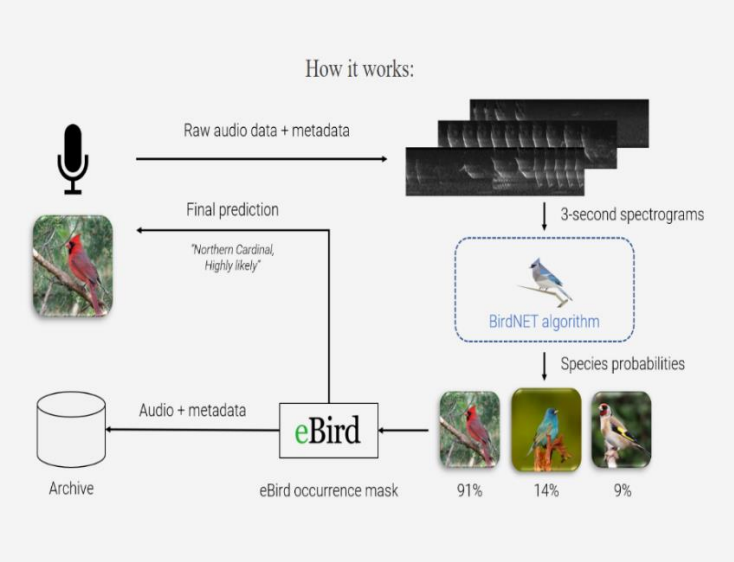

FIG 3.1 WORKIMG OF BIRD SOUND DETECTION

1) DATA COLLECTION

Acquire Bird Sound Recordings: Collect a diverse and representative dataset of bird sound recordings. You can use field recordings, online databases like Xeno-canto, or record sounds yourself using specialized equipment.

Metadata Annotation: Ensure each recording is associated with metadata such as species labels, recording location, date, and time. Accurate metadata is essential for training and evaluation.[10]

2) DATA PREPROCESSING

Audio Segmentation: Split the audio recordings into shorter segments (e.g., 1-5 seconds) containing individual bird vocalizations. This helps isolate and analyze specific bird calls or songs.

Feature Extraction: Extract relevant features from the audio segments. Commonly used features include:

a) Mel-frequency cepstral coefficients (MFCCs)

b) Chroma features

c) Spectral contrast

d) Zero-crossing rate

e) Energy

f) Pitch-related features

Data Augmentation: To increase the size and diversity of your dataset, you can apply data augmentation techniques such as pitch shifting, time stretching, and adding background noise to the audio segments.[11]

SPECTROGRAM GENERATION:- A spectrogram is a type of graphical representation that shows the variation in the frequency range with time. It typically consists of the x-axis range of frequency, the y-axis range of frequency, and the color of the representation to depict the intensity or power of that specific frequency. It can be generated by converting the signal in the time domain to frequency.[12]

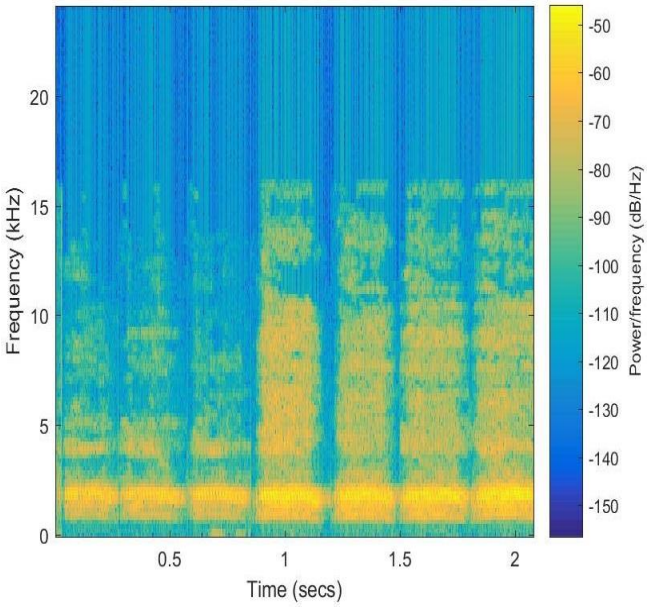

FIG.4 EXAMPLE OF A GENERATED SPECTROGRAM

3) MODEL SELECTION

Choose an Architecture: Decide on the machine learning model architecture to use. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or hybrid models like Convolutional Recurrent Neural Networks (CRNNs) are commonly used for bird sound recognition

Convolutional Neural Network:- A Convolutional Neural Network (CNN) is a type of artificial neural network designed specifically for processing structured grid-like data, such as images and 2D arrays. CNNs have revolutionized the field of computer vision and image processing and are widely used for tasks like image classification, object detection, facial recognition, and more.[13]

K-Nearest Neighbour Algorithm:- A K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for both classification and regression tasks. It's a type of instance-based learning or lazy learning, where the model stores the entire training dataset and makes predictions by finding the K nearest data points to a new, unseen data point based on a similarity measure.[14]

COMPARISION OF BETWEEN MODELS:-

1) KNN V/s CNN

Key Points	KNN	CNN
Algorithm Type	Instance-based algorithm	Neural network architecture
Training	Direct use of training data	Iterative training process
Complexity	Simple and low complexity	Complex due to layered architecture
Computation	Low during training, higher during prediction	High due to layered computations
Feature Extraction	Direct use of provided features	Learns hierarchical features from data
Pattern Recognition	Uses distance metrics to recognize patterns	Learns hierarchical features from data
Generalization	Simple generalization	Complex generalization to new data

TABLE 1.1 COMPARIOSION B/W KNN AND CNN

2) ANN V/s CNN

Key Points	ANN	CNN
Basics	One of the simplest type of the neural Network.	One of the most popular types of the Neural network
Data Type	Fed on tabular and text data	Relies on image data
Complexity	Simple in Constrast with the other two models	Complex due to layered architecture
Commendable Feature	Ability to work with incomplete Knowledge and high fault tolerance.	Accuracy in recognizing images.
Feature Type: Spatial Recognition	No	Yes
Uses	Complex problem solving such as predictive analysis.	Computer vision including image Recognition.

TABLE 1.2: COMPARIOSION B/W ANN AND CNN

3) RNN V/s CNN

Key Points	RNN	CNN
Basics	Most Advanced And Complex neural Network.	One of the most popular types of the Neural network
Data Type	Trained with sequences data.	Relies on image data
Complexity	Fewer Features than CNN but powerful due to Self-learning and Memory Potential.	Complex due to layered architecture
Commendable Feature	Memory and Self-learning.	Accuracy in recognizing images.
Feature Type: Spatial Recognition	No	Yes
Uses	Natural Language processing including sentiment analysis and speed recognition.	Computer vision including image Recognition.

TABLE 1.3: COMPARIOSION B/W RNN AND CNN

Transfer learning is a commonly used machine learning technique primarily used when the available dataset is limited. It is a process where the network is trained on a largedataset with similar data and then modifying the same to work well on the target data. In this particular application, we identify the bird species based on their Spectrogram images of the bird sounds. AlexNet was trained using transfer learning to recognize new categories of images which are thespectrogram images of the sounds of various bird species.[15]

4) MODEL TRAINING

a. Data Split: Divide your dataset into training, validation, and test sets (e.g., 70-15-15 split) to assess model performance accurately.

b. Loss Function: Select an appropriate loss function (e.g., categorical cross-entropy) for your classification task.

c. Optimization: Choose an optimization algorithm (e.g., Adam, SGD) and tune hyperparameters like learning rate and batch size.

d. Training: The training steps to improve the performance of your model. You can also implement various techniques such as early stopping to prevent the model from overfitting.

Number of species	Data Split	Epoch	Accuracy
2	80:20	20	92%
2	70:30	20	90%
4	80:20	20	88%
4	70:30	20	85.25%
4	80:20	35	97%
4	70:30	35	94%

TABLE.1.2: TRAINING RESULTS

5) MODEL EVALUATION

Test Set Evaluation: Assess your trained model's performance on the test set using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

Cross-Validation: Perform k-fold cross-validation to obtain a more robust estimate of your model's performance.

I. REAL-TIME IMPLEMENTATION

We can use the AlexNet neural network to predict the bird species that will appear in an input sound recording. Unfortunately, these predictions are not reliable since these recordings were made in environments with no noise.The AlexNet neural network can give a prediction based on a dataset that doesn't contain any noise. Unfortunately, in real-time recordings, the network encounters the issue of ambient noise, which can be caused by various factors such as vehicular sounds and overlapping human voices.Before implementing the network in real-time, it must first be thoroughly trained to ensure that it performs as well as it did in a simulation. This process should involve re-training the network on a dataset that is ideal for the model and sound samples that were collected from an environment that is not noisy..

The audio clips of various bird species are randomly chosen and converted into a fixed 48000 or 44100Hz sampling rate to preserve their diversity and prevent overfitting. The bit rates are also set to 320kbps and 128kbps to ensure that the recording quality is clear and has a low file size, ideal for audio applications.After converting the audio clips into bit streams and sampling rates, the spectrograms for each audio clip are generated. These are then used to retrain the neural network. After the learning process is complete, the model can be reused to classify the audio signalThe Neural Network can be retrained using these spectrograms. After transfer learning has been completed, the model can be reused to classify the audio signal. It is recommended to set the microphone's sampling rate to 44100 hertz and its bit rate to 128 kbps.

A real-time environment was used to test the system and produce classification results that were 91% accurate. The system was operated using a GUI, which involved executing various processes, starting from recording the sound to processing the data.

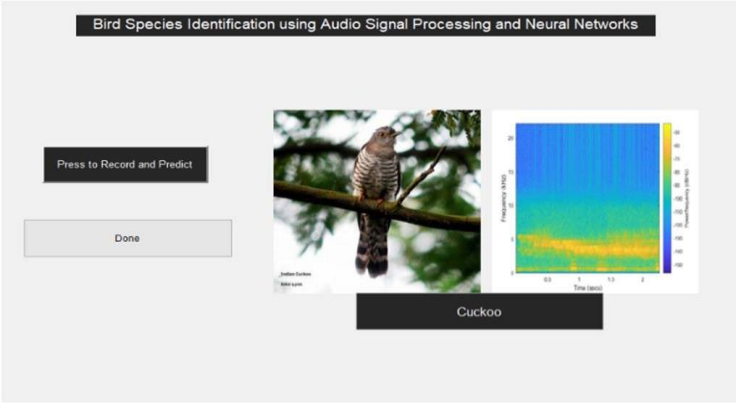

FIG1.5 REAL-TIME OUTPUT

II. FUTURE SCOPE

The goal of this project is to develop a system that can perform accurate and timely bird identification and prediction. It can be used on mobile devices. Through an app, users can record the bird sounds and then it will process the data and return the results.

This data will allow us to collect important information about birds, such as their movements across different areas and the number of species in a particular locality.[16]

III. CONCLUSION

In conclusion, looking into ways to predict and use technology to recognize bird sounds is really exciting. It can help us learn a lot more about how bird environments work and how birds behave. By using smart computer programs, special audio methods, and clever sensors, scientists and people who protect nature can make big changes in studying birds. When we can tell exactly which bird is making a sound, we can learn about how their homes are doing, where they travel, and how the types of birds are changing over time. But we need to understand that there are difficulties in this area. The different and sometimes complicated sounds that birds make make it hard to create strong computer programs that can recognize them well. Also, we have to think about what's right when using audio recording devices outside. We should be careful so that we don't bother the birds too much.

Even though there are problems, we've still made big progress in the last few years. As technology gets better and more people collect information, the computer systems that recognize bird sounds will probably get much better at their job. This will really help in protecting birds and keeping an eye on the environment. It will also help us learn more about how nature works in general.

REFERENCES

[1] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep Learning". Nature. 521 (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539. PMID 26017442. S2CID 3074096.

[2] later, Peter J. B.; Mann, Nigel I. (2004). "Why do the females of many bird species sing in the tropics?". Journal of Avian Biology. 35 (4): 289–294. doi:10.1111/j.0908-8857.2004.03392.x.

[3] "Bird Audio Detection challenge". *Machine Listening Lab at* Queen Mary University. *3 May 2016*. Retrieved 22 July 2018.

[4] "Watch out, birders: Artificial intelligence has learned to spot birds from their songs". *Science | AAAS*. 18 July 2018. Retrieved 22 July 2018.

[5] "Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation". *DeepLearning 0.1*. LISA Lab. Archived from the original on 28 December 2017. Retrieved 31 August 2013.

[6] Yang, Z.R.; Yang, Z. (2014). *Comprehensive Biomedical Physics*. Karolinska Institute, Stockholm, Sweden: Elsevier. p. 1. ISBN 978-0-444-53633-4. Archived from the original on 28 July 2022. Retrieved 28 July 2022.

[7] Cover, Thomas M.*;* Hart, Peter E. *(1967).* "Nearest neighbor pattern classification" (PDF). *IEEE Transactions on InformationTheory*.
***13** (1):2127*. CiteSeerX 10.1.1.68.2616. doi*:*10.1109/TIT.1967.1053964. S2CID 5246200

[8] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 2016, pp. 1–6.

[9] Venkatesan, Ragav; Li, Baoxin (2017-10-23). Convolutional Neural Networks in Visual Computing: A Concise Guide. CRC Press. ISBN 978-1-351-65032-8.

[10] Collobert, Ronan; Weston, Jason (2008-01-01). "A unified architecture for natural language processing". *Proceedings of the 25th international conference on Machine learning - ICML '08*. ICML '08. New York, NY, USA: ACM. pp. 160–167. doi:10.1145/1390156.1390177. ISBN 978-1-60558-205-4. S2CID 2617020

[11] Zhang, Wei (1988). "pattern recognition neural network and its optical architecture". Proceedings of Annual Conference of the Japan Society of Applied Physics.

[12] Sejdic, E.; Djurovic, I.; Stankovic, L. (August 2008). "Quantitative Performance Analysis of Scalogram as Instantaneous Frequency Estimator". *IEEE Transactions on SignalProcessing*. Bibcode:2008ITSP...56.3837S. doi:10.1109/TSP.2008.924856. ISSN 1053-587X. S2CID 16396084.

[13] "Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation". *DeepLearning 0.1*. LISA Lab. Archived from the original on 28 December 2017. Retrieved 31 August 2013.

[14] Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281 [cs.CL].

[15] West, Jeremy; Ventura, Dan; Warnick, Sean (2007). "Spring Research Presentation: A Theoretical Foundation for Inductive Transfer". Brigham Young University, College of Physical and Mathematical Sciences. Archived from the original on 2007-08-01. Retrieved 2007-08-05.

[16] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291–1303, 2017.