2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP) | 978-1-6654-3277-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCSP52374.2021.9465511

In-Vehicle Speech Command Operated Driver Assist System for Vehicle Actuators Control using Deep Learning Techniques

Prasanna Gawade
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.p2ael19015@cb.students.amrita.edu

Suresh Kumar P
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
p_sureshkumar@cb.amrita.edu

***Abstract*—Considering road accidents due to distracted driving, according to NHTSA around 10% to 17% injuries and fatalities are reported every year. Last year NHTSA logged 3142 fatalities due to distraction. To minimize this, considering driver's safety and convenience, a speech command operated in-vehicle actuators control driver assist system using Deep Neural Network is proposed. The proposed system assists to keep eye on the road, and hands on steering wheel ensuring safety. Existing actuation are manually performed by driver causing distraction thus developing an in-vehicle customized driver assist system for accepting speech command inputs, processing, recognizing, classifying using deep neural networks and actuating the desired vehicle actuator.Customized speech commands for Automotive Actuators Control (AAC) and generic benchmark Google Speech Commands (GSC v1) are validated using CNN and ConvLSTM model.Classification accuracies obtained are 94% and 98.30% for AAC and 94.47% , 92.20% for (GSC v1).**

***Index Terms*—Driver assist system, Customized speech commands recognition, Vehicle actuators control, Deep Neural Networks**

I. Introduction

Audio signal content classification points to the study of audio data stream for functional information. With respect to audio signals, the frame is precisely used for extracting and recognizing audio voice commands in signal which varies with time. The research intent is focused on speech pattern recognition for automotive actuators control applications with certainty that human voice involves categorization of strongly structured and organized audio samples. In the review of speech command recognition, Google, Amazon provides smart devices with voice assistance and keyword spotting systems on remote device allowing users to search using speech command [1]. Research has been done using Hidden-Markov Model (HMM), Hough-transform, matrix factorizing, and Radon transformation in area of audio signal classification which are complex in nature consisting of language, acoustic models. It requires linguistics models such as tokenization, dictionary of pronunciation and phonetic dependent trees [2].A simple light weight neural network model which is capable of running and predicting the class of the speech command in local-end processor is developed. Comparative analysis is done using 1) Convolutional Neural Networks (VGG16) 2) ConvLSTM for Google Speech Command version-1 (GSC v1) and Automotive Actuators Command Dataset (AACD).

Identification of speech commands which includes phrases i.e. limited word-audio clips has variety of applications and have gained significant market area with the wide usage of voice operated embedded systems. Along-side, it spans to development of an Open Keyword Search system, content extraction in phrased dialogues, specific word identification in music-voice data-base cataloging, audio monitoring, and natural sound classification.

Existing networks are computationally costlier compared to limited vocabulary speech command detection models which are generic in nature and require large number of data words (vocabulary) to train [3].Motivated by the improvements in classification results, our research investigates the adaptability and agility of deep Convolution Neural Networks for Automotive Actuators Commands Dataset.

II. LITERATURE SURVEY

Researchers in recent years, have proposed various techniques to obtain accurate voice audio signal recognition and classification. To conquer the primitive difficulty of geographical accent variations researches are conducted using methods like Hidden-Markov Model, Hough transformation, matric factorizing, Radon transformation, RBMs and Deep NNs in the field of voice command classification.

Speech command include various semantic and para- linguistic features such as accent, age and emotional state of speaker [2] [4].The BCNN is effectively classifying Google speech command dataset (v1) and US8K by imitating spatial repre-

978-1-6654-3277-1/21/$31.00 ©2021 IEEE

sentations by extracting features for Spectrogram as perceived by receptive neurons by auditory cortex of mammals [5]. A detailed study was done on Deep Neural Networks by H.Meng et al. [6] demonstrated that for image classification with 3D Mel spectrogram for Speech emotion recognition are best suited as spectrograms represent spatial transformations in the form of images with respect to time [7] [8]. CNN has be trained by Karol [9] for Environmental Sound Classification-10 (ESC-10) dataset, and results show robust performance for systems with noise represented in the form of spectrograms. Existing networks are computationally costlier compared to limited vocabulary speech command detection models which are generic in nature and require large number of data words (vocabulary) to train. Zero crossing rate is one of the important aspect in speech recognition in order to detect speech activity and is computed by method of auto correlation [10],features considered during detection of background applause in speech signal is studied which extracts auto correlation,energy decay factor and the MFCC (Mel Frequency Cepstral Coefficients) [5]. Oliveira et al. [11] extracted patterns of binary locale of audio from frequency–time spectrogram. The para-linguistic features of audio speech command vary and depend on various factors such as, environment in which it was recorded, type of transducers used, sample rate and amplitude of audio, noise presence during data collection in real–time. All these contribute in overall system classification accuracy and are considered vital during feature extraction [5] [9] [12]. In the past years researches have mainly concentrated on describing audio voice signal by visual formed by Mel-spectrogram for image classification [5]. Costa et al. [11] featured graphics from pre-defined window of a Mel-spectrogram obtained from Mel-scale windowing with collaboration of Support Vector Machines classifiers. The subsequent work they represented combined acoustic visual features, which improved the classification metric .Having said that, instead of relying on an approach which varies with time, the audio voice command was featured as a patch of visual graphics on the basis of Mel-Spectrogram.

III. DATASET AND METHODOLOGY

A. Dataset generation

One of the dataset considered is Google Speech Command(v1) which consists of 64272 short audio clips of single English word with audio length of 1s. Out of 30 classes the goal is to classify 'right' 'left' 'up' 'down' 'yes' 'no' 'on' 'off' 'stop' 'go' and remaining 20 classes are considered as Unknown.

Our synthesized dataset consists of Automotive Actuator Control commands (AACD) , consist of 2.5s audio samples with 30 actuators to control in a vehicle such as ' Open Sunroof' 'Close sunroof' 'Headlights on' Headlights off' 'AC on' 'AC off' 'open right-mirror' 'close-right mirror' and 'open left-mirror' 'close left -mirror' and such 20 other commands are generated. The parameters considered during collection and synthesizing AACD were para-linguistic feature of each command [4]. Utilizing online platform Play.ht which synthesizes audio data using TTS(Text to Speech) of around 260+ voices from Google , Amazon, Microsoft and IBM, the audio signals have been generated for each of the commands considering dialects across the globe (with more concentration on Indian subcontinent –Hindi, Marathi ,Telugu ,Bengali accent) with equal distribution of Male and female voices. Total of 100 (50-Male, 50-Female) speakers from synthesized generators and 10 (5-Male, 5-Female) real-time recorded audio data samples are collected.The sample rate was set to 44.1KHz for generation and recorded with mono channel,stored in uncompressed data format and total of 110*30 raw audio commands are generated. Neural Networks perform more efficiently and accurately when trained on large scale feature extracted data to diversify data and increase the variance of dataset, audio augmentation techniques are implemented using Librosa (python library for visualization and recognition of speech and audio data) and Audacity-open source software for audio related operations. Change in pitch, change in tempo, change in speed, overlaying of standard and application specific(Automotive environment) noises such as (in-cabin noise, traffic noise-Type1,Type2,Type3) with relatively less amplitude as compared to raw audio input,all these audio augmentation techniques are implemented along with raw samples to add variation and diversity in AACD. Total of 33000 samples for 30 classes with 1100 audio samples for each class collectively form dataset. Experimental analysis is done for 10 classes with data split for 80% for training and 20% validation.

Application of audio augmentation on AA commands Dataset is performed and it is observed that change in spatial and temporal parameters will lead to more diverse and variation in dataset thus improving overall accuracy and avoid over-fitting due to less number of samples in dataset. Details are tabulated in TABLE I.

TABLE I: AUDIO AUGMENTATION TECHNIQUES.

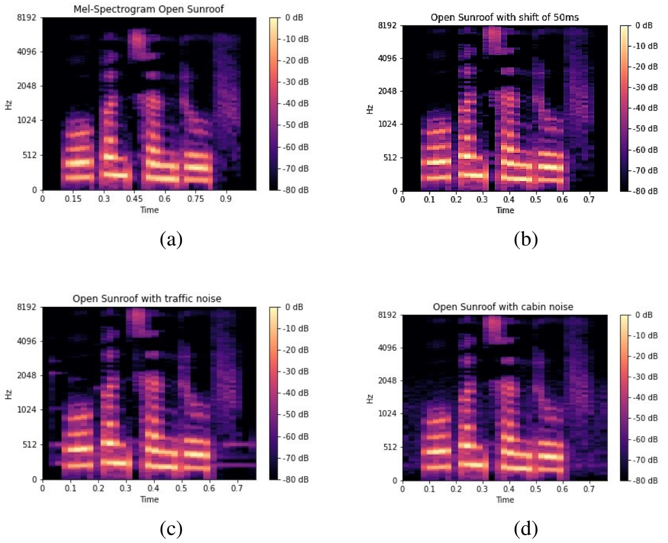
Factors	**Levels**	**Effects on audio -Responses**
Change in Pitch	- (20%) to + (20%) step size ± 5%	Male to Female voice conversion and vice-versa
Change in Speed	- (20%) to + (20%) step size ± 5%	Variations in rate of speech delivery
Change in Tempo	- (20%) to + (20%) step size ± 5%	Variations w.r.t contextual and emotional factors

The Mel-Spectrograms in Figure 1 show the effects of audio augmentation on 'Open sunroof' command.

B. Methodology

In this paper,we have explored behavior of CNN and ConvLSTM for GSC-v1 and customized dataset (AACD) consisting of 2 word, 3 word phrases generated in stationary/dynamic noise environment [13]. The proposed schematic system is represented in Figure 2.

- Audio Preprocessing:The audio commands vary in length based on the number of words in it. In order to avoid biasing effect in learning algorithm , necessary audio



(a) (b)

(c) (d)

Fig. 1: Spectrogram generation of (a) Raw Open Sunroof (b) Shift of 50ms (c) Overlaid traffic noise (d) Overlaid in-cabin noise

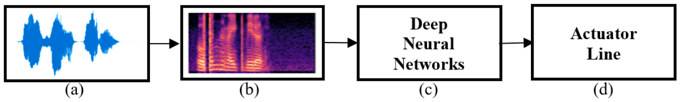


Fig. 2: Schematic flow diagram of system
(a) Audio input and preprocessing, (b) MFCC/ Mel Spectrogram extraction and generation, (c) Deep Neural Networks, Softmax layer probabilities of output being correct (d) output signal mapping and selection.

files are padded with zero-valued samples (silence) to maintain uniformity throughout the dataset.

- 2D Spectrogram Generation:Raw audio files are transformed into two dimensional visual images represented as spectrograms [14]. It plots change in frequency with respect to time in a 2-D feature map, as human ear perceive change in frequency in logarithmic scale, audio frequencies are mapped on to log scale thus converting spectrogram into Mel-scale spectrogram.Transformation FFT of 1024 points is used with 50% overlap to ensure regularity in speech transformed samples. The obtained spectrogram is resized to (224*224) and fed as input to model for classification.
- CNN architecture:The standard input layer of CNN accepts a input image with $I \in \mathbb{R}^{(s*s*c)}$,where c and s are channels and image size at the input,respectively.A filter is defined and shared across the image input with previous and subsequent layers extracting local features with translational invariance, these layers are preceded by pooling and subsampling which abstract relevant features while minimising spatial resolution.Filters defined learn relevant features for accurate spectrogram classification.Optimisers used learn suitable kernels effectively. Window of moving average is utilized by Adagrad and adapts to learning rate [5].Equation of squared gradients is in equation 1.

$$\overline{g^2_{MA}} = \frac{g^2_m + g^2_m + ... + g^2_{m-(n-1)}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} g^2_m - i \quad (1)$$

For every new instance, the ultimate layer is Softmax function, which outputs probability distribution over 10 classes.The audio classification problem is basically solved as image classification.
- RNN architecture: Considering a time varying system over spatial domain presented by $N \times M$ where N rows and M columns in a cell which consists O measurements over time.The data feature at any time can be extracted from tensor $K \in PO \times N \times M$, where P represents region of feature.
Spatio-temporal data is not handled effectively by fully connected LSTM [15] [16]. Convolution neural networks helps in retrieving information of spatial features using its tensors. We adopted convolution based LSTM model ConvLSTM which predicts the future state value by taking past state and current state into consideration. Kernels in ConvLSTM plays an important role to deal with small and large change in data. This property plays an important role in classifying time varying speech signals. Similarly, we are using zero padding which can help in distinguishing continuous speech signals efficiently.
ConvLSTM follows folding – unfolding structure to predict the output same as input. It resembles encoder decoder architecture Folding LSTM $f_{encoding}()$ converts higher dimensional sequences into lower dimension while unfolding structure is just like decoder $g_{decoding}()$helps us in getting information from same as that of input.X_N and Y_N are input and output sequences respectively in equation 2.

$$Y_N = g_{decoding}(f_{(encoding)}(X_N)) \quad (2)$$

On a single image feature, convolution operation is performed by layers and resultant features are flattened to 1D array, this operation is performed over all images in time frame ,which is fed as input to LSTM layer. We followed ConvLSTM architecture to get a robust model which can help in classifying diverse speech signals as described in TABLE II.
The Mel- Frequency Cepstrum Co-efficient are extracted and fed as input tensor to conv2d layer of ConvLSTM, sub-sampling is preferred at necessary stages to preserve spatial and temporal relation within layer transitions. Convolutional layer (kernels) perform feature extraction and LSTM layers update weights based on forward and backward transition method.

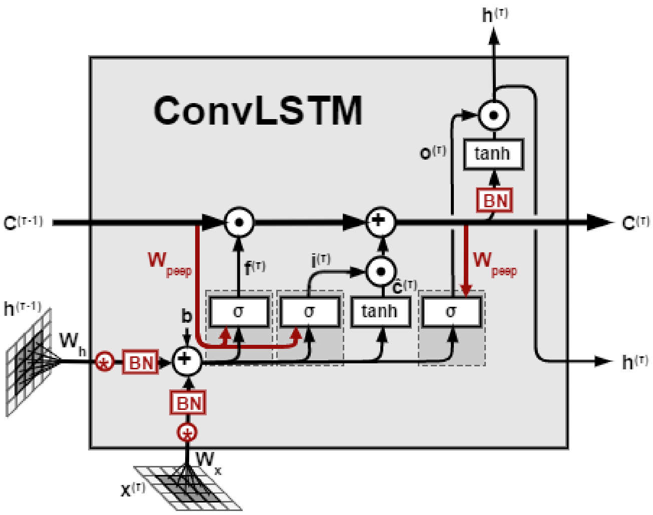


Fig. 3: A convLSTM cell

IV. Experimental Setup

The part of audio data samples are collected and stored in lossless uncompressed format which are recorded using omni directional microphone with and without noise cancellation having 360º audio pickup coverage, in the presence of ambient environmental noise with sample rate of 44.1KHz and in mono channel mode. It is ensured that samples are uniform in length. Use of Librosa and Audacity is done for preprocessing of audio files on i5 6th Gen Intel-processor with 8GB RAM operated at 2.30GHz. Setup for training models is done using Google Colab GPU- Tesla T4 TU104 for both CNN and ConvLSTM having 16GB RAM. Training time varies based on hardware assigned and used. Tensor board package from Tensorflow is used for analysis of metrics. Pre-recorded and synthesized samples are considered for testing the performance of model.
Training on both Google Speech Command dataset (v1) having one word utterances and our AAC dataset is validated and tested with test dataset. Following metrics are split of dataset with 80% training and 20% validation.

V. Result and Analysis

The accuracies are noted and represented using chart for better comparative analysis in Figure 4. It shows the model accuracies for defined datasets. Automotive Actuators Control Dataset (AACD) and generic Google speech command dataset (GSC v1) is validated using CNN and ConvLSTM model. Resulted in classification accuracy of 94% and 98.30% for AACD and 94.47%, 92.20% for (GSC v1).

It is observed that CNN and ConvLSTM performs well for GSC(v1), ConvLSTM evaluation on sparse categorical accuracy shows good results ,and tends to over-fit resulting in non-generalization of model. This can be avoided by adding diverse raw samples into dataset. CNN model performance degrades when noise is dominant over speech commands, this can be eliminated by recording data with noise cancellation and high fidelity microphones .On benchmark dataset performance can be further generalized by adding more background noise samples into the data.

TABLE II: Model Summary

Layer (type)	Output Shape	Param #
conv2d	(None,88,128,50)	300
batch_norm	(None,88,128,50)	200
max_pooling2d	(None,44,128,50)	0
dropout	(None,44,128,50)	0
conv2d_1	(None,44,128,60)	27060
batch_norm_1	(None,44,128,60)	240
max_pooling2d_1	(None,22,64,60)	0
dropout_1	(None,22,64,60)	0
conv2d_2	(None,22,64,80)	43280
batch_norm_2	(None,22,64,80)	320
max_pooling2d_2	(None,11,32,80)	0
dropout_2	(None,11,32,80)	0
reshape	(None,352,80)	0
bidirectional	(None,352,128)	74752
bidirectional_1	((None,64))	41472
dense	(None,50)	3250
dense_1	(None,20)	1020
dropout_3	(None,20)	0
dense_2	(None,10)	210

Total params: 192,104
Trainable params: 191,724
Non-trainable params: 380

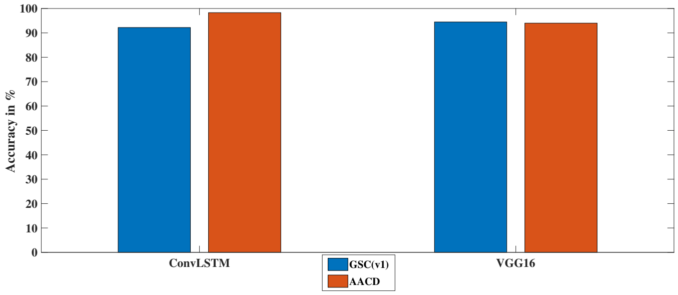


Fig. 4: Comparison chart of NN on AACD and GSC (v1)

One real-time sample from new user is taken to perform test for ‘Close left mirror’ command with similar pre-processing stage at input the output results were satisfactory with 0.65 probability for desired class and probability of 0.34 for ‘Open left mirror' , and remaining 0.1 is distributed across 8 classes. It is observed that model is prone to change in input audio w.r.t noise dominance over threshold. When tested with low amplitude noise overlaid on to audio signal, taken from unseen speaker, model classifies it accurately with poor repeatability and reproducibility.

The testing is performed on synthesized sample of 'Close left mirror' and model classifies it with probability (0.98) being audio input as 'Close left mirror'. Output predicted probability plot for synthesized 'Close left mirror' is shown in Figure 5 Around 5% of the total dataset i.e. up to 750 samples (75 samples) per class are classified, All the samples used for testing belong to sythesized dataset and no real time samples were used. testing metrics considered is confusion matrix. Matrix plot is in Figure 6. The model outputs are considered for top-k predictions, with k=3 for top-k probabilities being used to query the command in-case if command is unclear. Based on the value of 'k' the command query can be prompted to select correct course of action.

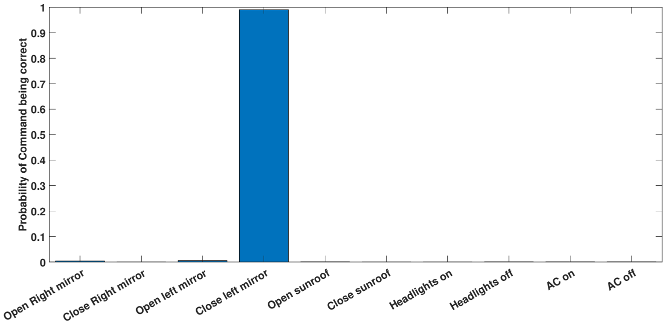
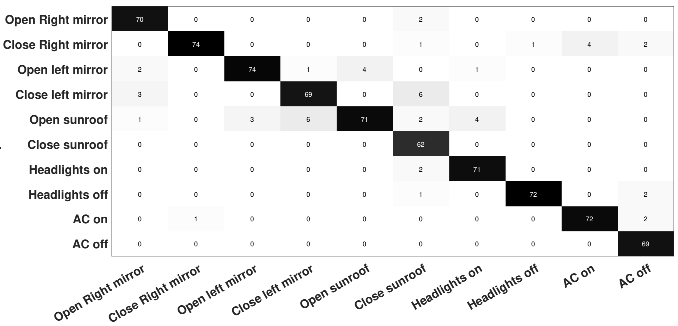


Fig. 5: Output layer probability distribution of VGG16 for Close left mirror



	Open Right mirror	Close Right mirror	Open left mirror	Close left mirror	Open sunroof	Close sunroof	Headlights on	Headlights off	AC on	AC off
Open Right mirror	70	0	0	0	0	2	0	0	0	0
Close Right mirror	0	74	0	0	0	1	0	1	4	2
Open left mirror	2	0	74	1	4	0	1	0	0	0
Close left mirror	3	0	0	69	0	6	0	0	0	0
Open sunroof	1	0	3	6	71	2	4	0	0	0
Close sunroof	0	0	0	0	0	62	0	0	0	0
Headlights on	0	0	0	0	0	2	71	0	0	0
Headlights off	0	0	0	0	0	1	0	72	0	2
AC on	0	1	0	0	0	0	0	0	72	2
AC off	0	0	0	0	0	0	0	0	0	69

Fig. 6: Confusion matrix for Spectrogram classification of 10 classes

The model is tested on diverse test samples and meets expectations in the presence low amplitude background noise.

VI. Conclusion

Experimental results demonstrate that ConvLSTM model is validated and CNN is effectively classifying the inputs into desired output with minimum computational cost and complexity compared to online speech detections methods and process relatively faster in offline mode, methods have to be incorporated to improve repeatability and reproducibility. More audio samples can be added to make models more robust thus achieving improved performance. The models can be deployed into automotive environment for which it was developed with necessary pre-processing techniques. Overfitting of data is observed, since all audio samples in a class resemble similar audio output pattern. We conclude that model accuracy and efficiency can be improved by incorporating more diverse samples avoiding over-fitting.

Future work to be carried out includes more number of actuators to be controlled within vehicle. All the simulation and testing to be tested with real time hardware implementation in a vehicle and testing its behavior on unseen speaker command, Speech processing and classification can be done using Jetson TX1 in collaboration with hardware in loop (HIL) test bench setup using dSpace-Micro Auto box II which probes into in-vehicle CAN (Controlled Area Network) for easy access of vehicle actuators.

Acknowledgment

We acknowledge time and active participation of my colleagues for contributing towards audio speech command dataset.

Special thanks to Play.ht for providing a platform for Data synthesizing with diverse dialects.

References

[1] Lalitha, S., Tripathi, S. Gupta, D. Enhanced speech emotion detection using deep neural networks. Int J Speech Technol 22, 497–510 (2019). https://doi.org/10.1007/s10772-018-09572-8

[2] Watanabe, Shinji, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. "Hybrid CTC/attention architecture for end-to-end speech recognition." IEEE Journal of Selected Topics in Signal Processing 11, no. 8 (2017): 1240-1253.

[3] Sainath, Tara N., Brian Kingsbury, George Saon, Hagen Soltau, Abdelrahman Mohamed, George Dahl, and Bhuvana Ramabhadran. "Deep convolutional neural networks for large-scale speech tasks." Neural networks 64 (2015): 39-48.

[4] Alkhawaldeh, Rami S. "Dgr: Gender recognition of human speech using one-dimensional conventional neural network." Scientific Programming 2019 (2019).

[5] Sinha, Harsh, Vinayak Awasthi, and Pawan K. Ajmera. "Audio classification using braided convolutional neural networks." IET Signal Processing 14, no. 7 (2020): 448-454.

[6] Meng, Hao, Tianhao Yan, Fei Yuan, and Hongwei Wei. "Speech emotion recognition from 3D log-mel spectrograms with deep learning network." IEEE access 7 (2019): 125868-125881

[7] Srinivasan, Sriram, Vinayakumar Ravi, V. Sowmya, Moez Krichen, Dhouha Ben Noureddine, Shashank Anivilla, and Soman Kp. "Deep convolutional neural network based image spam classification." In 2020 6th conference on data science and machine learning applications (CDMA), pp. 112-117. IEEE, 2020.

[8] Sasidhar, T. Tulasi, K. Sreelakshmi, M. T. Vyshnav, V. Sowmya, and K. P. Soman. "Land cover satellite image classification using ndvi and simplecnn." In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2019.

[9] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6. IEEE, 2015.

[10] Kathirvel, P., M. Sabarimalai Manikandan, S. Senthilkumar, and K. P. Soman. "Noise robust zerocrossing rate computation for audio signal classification." In 3rd International Conference on Trendz in Information Sciences Computing (TISC2011), pp. 65-69. IEEE, 2011.

[11] Costa, Yandre MG, L. S. Oliveira, Alessandro L. Koerich, Fabien Gouyon, and Jefferson G. Martins. ”Music genre classification using LBP textural features.” Signal Processing 92, no. 11 (2012): 2723-2737.

[12] Balasingam, M. D., and C. Santhosh Kumar. ”Refining Cosine Distance Features for Robust Speaker Verification.” In 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 0152-0155. IEEE, 2018.

[13] Kaushik, M., P. Prakash, R. Ajay, and S. Veni. ”Tomato Leaf Disease Detection using Convolutional Neural Network with Data Augmentation.” In 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1125-1132. IEEE, 2020.

[14] Shrawankar, Urmila, and Vilas M. Thakare. ”Techniques for feature extraction in speech recognition system: A comparative study.” arXiv preprint arXiv:1305.1145 (2013).

[15] Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. ”Convolutional LSTM network: A machine learning approach for precipitation nowcasting.” arXiv preprint arXiv:1506.04214 (2015).

[16] de Andrade, Douglas Coimbra, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. ”A neural attention model for speech command recognition.” arXiv preprint arXiv:1808.08929 (2018).