Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant

Fatma Sh. El-metwally [1], Ali I. Eldesouky [1], Nahla B. Abdel-Hamid [1], and Sally M. Elghamrawy [2,3,*]

[1] Department of Computer Engineering and Control Systems, Faculty of Engineering, Mansoura University, Mansoura, Egypt

[2] Department of Computer Engineering, MISR Higher Institute for Engineering and Technology, Mansoura, Egypt

[3] Scientific Research Group in Egypt (SRGE), Egypt

*Correspondence: sally@mans.edu.eg, sally_elghamrawy@ieee.org (S.M.E.)

***Abstract*—A virtual assistant has a huge impact on business and an organizations development. It can be used to manage customer relations and deal with received queries, automatically reply to e-mails and phone calls.Audio signal processing has become increasingly popular since the development of virtual assistants. Deep learning and audio signal processing advancements have dramatically enhanced audio tagging. Audio Tagging (AT) is a challenge that requires eliciting descriptive labels from audio clips. This study proposes an Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant to categorize and analyze audio tagging. Each input signal is used to extract the various audio tagging features. The extracted features are input into a neural network to carry out a multi-label classification for the predicted tags. Optimization techniques are used to improve the quality of the model fit for neural networks. To test the efficiency of the framework, four comparison experiments have been conducted between it and some of the others. From these results, it was concluded that this framework is better than the others in terms of efficiency. When the neural network was trained, Mel-Frequency Cepstral Coefficient (MFCC) features with Adamax achieved the best results with 93% accuracy and a 0.17% loss. When evaluating the performance of the model for seven labels, it achieved an average of precision 0.952, recall 0.952, F-score 0.951, accuracy 0.983, and an equal error rate of 0.015 in the evaluation set compared to the provided Detection and Classification of Acoustic Scenes and Events (DSCASE) baseline where he achieved and accuracy of 72.5% and a 0.209 equal error rate.**

***Keywords*—audio tagging, Deep Neural Networks (DNNs), optimizations, Detection and Classification of Acoustic Scenes and Events (DCASE)**

I. Introduction

Audio signal processing is the operation of applying powerful methods and techniques to audio signals [1].

Devices such as smartphones have been increasingly popular in recent years, and communicating remotely via the internet has become the preferred way to connect over face-to-facemeetings. However, auditory noise, distortion, and echo are unavoidable in any communication process. Every day, millions of multimedia recordings are created and published on the Internet as a result of the widespread use of electronic communication. These recordings contain a variety of media, including music, news broadcasts, television shows, and science articles. Recently, audio analysis has attracted a lot of interest from researchers. In reality, audio rarely comes from a single source but rather from a blend of sounds from various sources. Therefore, audio pattern recognition is a critical research topic in the field of machine learning and performs a significant role in our daily lives, e.g., for automatic audio tagging [2], audio segmentation [3], and audio context classification [4, 5].

The goal of Audio Tagging (AT) is to label a sound clip with one or more tags. “Tags” is the sound events that happen in the audio recording., such as “speaking”, “TV sound”, “Clapping”, “car”, and others. Audio labelling technology can be utilized in a variety of applications including lifelogging [6], medical activity surveillance [7], and so on.

Early, Special data sets collected by individual researchers were used for audio pattern recognition [8, 9]. For instance, Woodard [8] has implemented a Hidden Markov Model (HMM) to categorize three kinds of sounds: The wooden door was opened and closed, metal was dropped, and water was poured. Recently, the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series [10–13] has presented publicly available datasets, such as acoustic scene classification and sound event detection datasets.

For the musical tagging task, deep learning methods have proved their efficiency [14, 15]. Deep learning-based techniques have also been used for environmental audio tagging, it is a suggested task in the DCASE 2016 challenge [16] based on the CHiME-home dataset [17].

Until now, most of the audio-related recognition systems have been used, features are extracted from the frequency domain of the audio signal [18], such as Mel Frequency Cepstral coefficients (MFCCs) [19], log-frequency filter banks [20], and time-frequency filters [21].

The task of audio tagging has been extensively studied. The Gaussian Mixture Model (GMM) is trained on MFCCs [22], Convolutional Neural Networks (CNNs)

Manuscript received July 1, 2022; revised August 12, 2022; accepted November 11, 2022; published June 16, 2023.

with input features Constant-Q-Transform [23], and deep neural networks with inputs from the Mel Filter Bank [24].

This study presents an Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant for audio signal classification. First the features are extracted from the input signal, and then these features are input into a neural network for classifier. Adamax optimizer is used to optimize the parameters of the neural network. This study is organized as follows: Section II discusses related work, while Section III discusses the proposed Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant. Section IV discusses the experimental results. Finally, the conclusion and future work are presented in Section V.

II. Related Work

The DCASE challenge is a set of tasks designed to improve sound classification and detection systems.

The audio files for the challenge have two limitations that make the tagging task difficult. The first is environmental noise, which occurs because the recordings are taken in real-world environments. Therefore, it is necessary to choose a strong environmental noise resistance model for audio tagging.

The second limitation is that multiple sound sources can exist in a single recording. As a result, the classification model must be able to model and recognize multiple sound sources at the same time.

So, we must choose the optimal feature that is related to the problem and will solve it, the best deep learning model for classifier and optimizer, to improve efficiency, reduce resource consumption, and reduce time. Previously for the audio tagging task. The GMM was trained on audio features such as the DSCASE 2016 base line [22, 24], where GMM was trained on Mel frequency correlation coefficients and achieved an accuracy of 72.5% and a 0.213 equal error rate on the evaluation set. However, neural networks such as Recurrent Neural Network (RNN), Deep Neural Network (DNN), and CNN were trained on audio features.

Lars and Phan *et al.* [25] performed multi-label classification for audio tagging, a short-time Fourier transform was used to extract features from audio signals, which were then fed into a convolutional neural network with masked global pooling. They achieved an overall accuracy of 84.5% and an Equal Error Rate (EER) of 0.17 on average.

In Xu and Huang *et al.* [26], the Mel-Frequency Cepstral Coefficients and mel filter bank features were employed, and these were fed into a deep neural network with an SGD optimizer. The MFCC Feature produced results with an EER of 0.168 on average. The MFB Feature produced results with an EER of .157 on average. It turns out that the MFB feature outperforms the MFCC feature, so it took the MFB feature and tested it with another network that has a Denosing Autoencoder (DAE) with SGD optimizer and achieved results with an average EER of 0.148.

In Kong *et al.* [27], for feature extraction, 40 Mel-filter bank features are used. These features are input to Deep neural network with SGD optimizer for Classification. They achieved an overall EER of 0.209.

In Vu and Wang [28], a MFCCs input feature signal was used, each audio chunk was preprocessed by segmenting it with an .04-ms sliding window with a hop size of .02 ms and converting it to 13-dimensional MFCCs. The RNN system achieved a 0.200 EER.

In Xu and Kong [29], a Convolutional Gated Recurrent Neural Network (CGRNN), which is a hybrid of the CNN and the gated recurrent unit, was used. Three features, such as MFBs and spectrograms, as well as raw waveforms, were used to extract features from audio, and these features were fed into the CGRNN. To classify audio tagging, spatial features such as Interaural Phase Differences (IPD) or Interaural Time Differences (ITD) and Interaural Level Differences (ILD) are incorporated in the hidden layer. It looks like IMD has some meaningful patterns, whereas the ILD and IPD appear to be random, which would exacerbate the classifier's training difficulties. The spectogram achieved 0.110. The spectogram with the IMDs can get the EER of 0.104. The raw waveforms achieved 0.127. The raw waveforms with the IMDs can get the EER of 0.106. The MFBs achieved 0.119 and MFBs with the IMDs could procure the bare minimum of EER, which is 0.102.

III. The Proposed Optimized Deep Neural Network Audio Tagging Framework for Virtual Business

The proposed Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant consists of three layer as follows: The Data preparation layer, data modeling layer, and prediction layer (see Fig. 1).

A. Data Preparation Layer

1) Audio signal representations

The extraction of audio features is an important process in audio signal processing, which is a subcategory of signal processing. It deals with audio signal processing. By transforming digital and analogue signals, it eliminates background noise and balances the time-frequency ranges. It concentrates on arithmetic methods for changing sounds. Each audio signal contains numerous characteristics or features [30]. So, we should extract the features that are pertinent to the issue we are trying to solve.

An audio signal is a signal containing information in the audible frequency range. Audio is produced by the shaking of a body, and that shaking causes the wiggle of air particles, which leads to a change in air pressure. The combination of high and low air pressure causes a wave, and we can represent this wave using wave form, as shown Fig. 2.

A Fourier transform is used to decompose complex periodic audio into a sum of sine waves oscillating at different frequencies.

The Fast Fourier Transform (FFT) is a significant analytical technique in the field of audio. It decomposes a signal into its spectroscopic system and offers information about the signal's frequency.

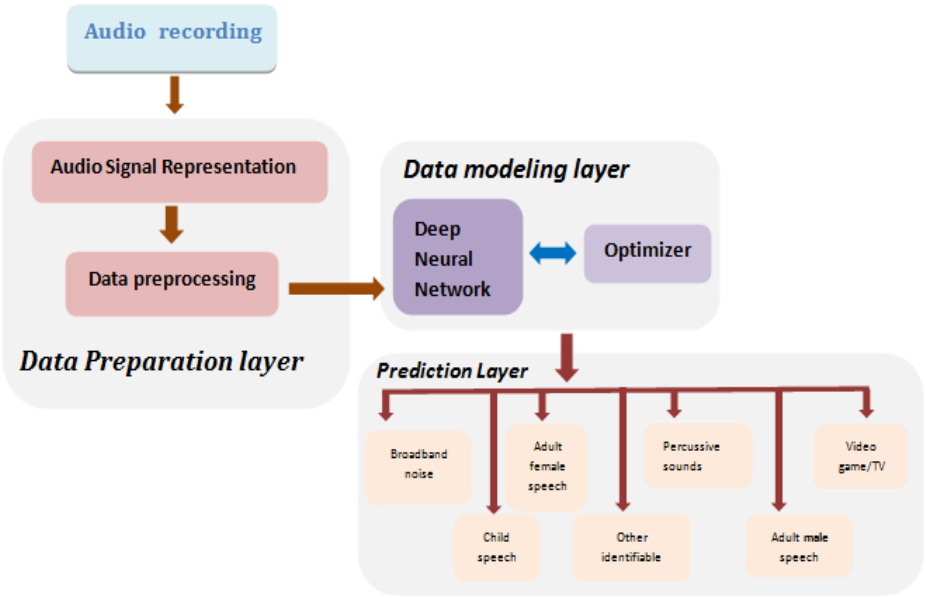


Figure 1. The proposed optimized deep neural networks audio tagging framework for virtual business assistant.

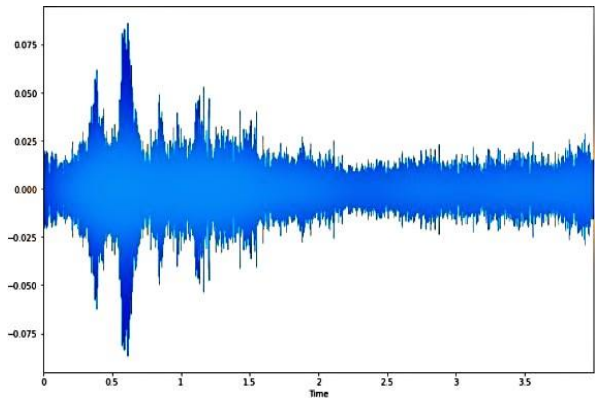


Figure 2. Plot audio wave in time domain.

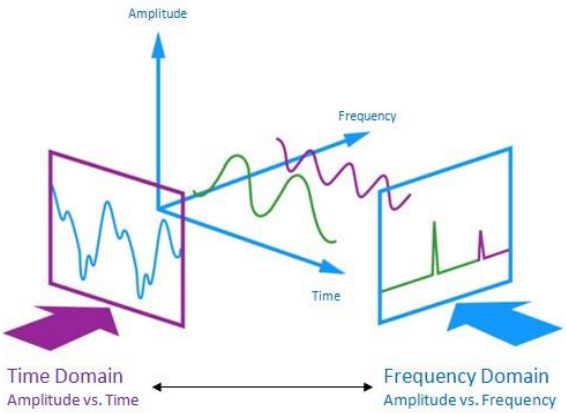


Figure 3. Plot FFT of the audio.

Note that when we do Fourier transform basically we move from the time domain to the frequency domain, as shown Fig. 3 and because of it we lose information about time. at first, it seems we lost a lot of information. but there is a solution to that, and it's called spectrogram, as shown in Fig. 4. Another feature that is fundamental and as important as spectrogram for deep learning, it called Mel Frequency Cepstral Coefficient (MFCC), as shown in Fig. 5.

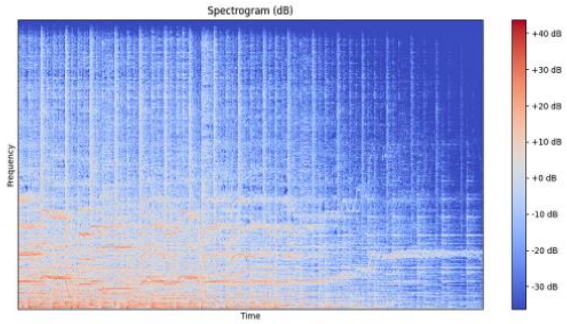


Figure 4. Plot STFT of the audio.

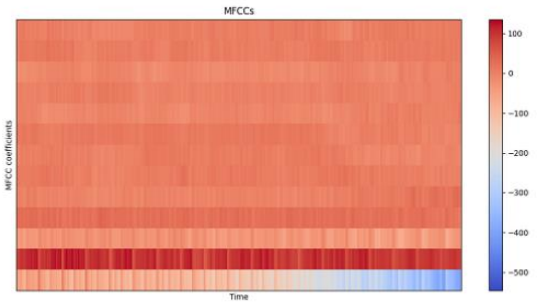


Figure 5. Plot MFCCs of the audio.

2) Data preprocessing

Original data sometimes contains noise, is incomplete, or is in an unsuitable format that makes it difficult to use it directly in deep learning models. Data preprocessing is one of the missions essential to purifying the data, which makes it fit for a deep learning model. Data preprocessing refers to the techniques that must be followed to transform or encode data in order for it to be conveniently analyzed by a machine.

Encoding Categorical data is information that has distinct categories within a data set. The machine learning model is based entirely on mathematics and numbers. However, if our dataset included a categorical variable, then it may cause issues during the model's construction. So, these category variables should be encoded as numbers [31].

Feature scaling is a method for standardizing independent features in a dataset on the same scale. The feature selection helps to perform calculations in algorithms extremely rapidly. It is a significant element in data preprocessing. A standardization technique is used. It is a very efficient technique that rescales the feature value so that it has a distribution of 0 mean and variance equal to 1 [32].

Here's the formula for standardization value:

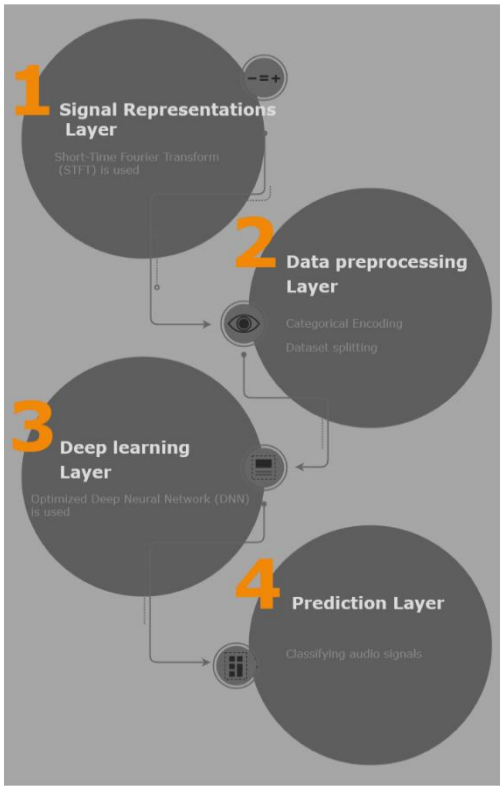
$$X_{new} = \frac{x_{i} - x_{mean}}{\text{standard deviation}} \quad (1)$$

Data set splitting: a data set splitting strategy is required for building a model with good generalization performance.

B. The Data Modeling Layer

Deep Neural Network (DNN) is a non-linear multi-layer model that can be used for classification [33] or regression [34] task. In the case of our audio classification issue, the input denotes the chain of audio features [35, 36], such as MFCC and STFT.

A DNN structure consists of three layers: input, hidden, and output. Fig. 6 shows the DNN structure. The hidden layer may have two or more layers, whereas the input and output layers are single layers. The hidden layer contains a set of neurons. The parameters used in a neural network are shown in Table I. The input layer receives data features. After processing in the hidden layers, prediction values are produced from the output layer.



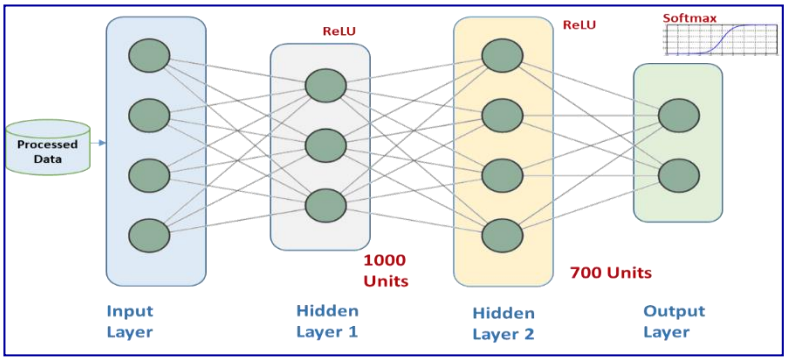


Figure 6. The deep neural network structure used in deep learning layer of ODLAT framework.

TABLE I. PARAMETERS OF THE NEURAL NETWORK

NN Parameter	Values
Classifier	Sequential
No of Hidden layer	2
No of Neuron in first hidden layer	1000
No of Neuron in second hidden layer	700
Hidden activation function	ReLU
Output activation function	Softmax
Optimizer	Adamax
loss function	Categorical cross entropy
Batch Size	100
No of Epoch s	100
Learning rate	0.005
momentum	0.9

Neural networks are prone to over-fitting as a result of the vast number of variables. Dropout is a regularization method that avoids over-fitting in neural networks. Dropout denotes the elimination of all incoming and outgoing connections as well as all hidden and visible units from a neural network. Each neuron in a neural network is eliminated with a probability of 0.5 during each training iteration of the original approach, and all neurons are included during testing [37].

C. Optimizer

During the training of the deep learning model, we have the function of loss. which tells us about the weakness of the model at the moment. So, we must utilize this loss to

train our network to do well. Basically, what we have to do is use the loss and try to reduce it.

Because reducing the loss makes the model work better. An optimizer's primary function is to change the neural network's parameters for example weights and learning rate. As a consequence of this, it aids in lowering overall loss and improving accuracy [38]. The Adamax optimizer is employed in this study.

D. The Prediction Layer

This layer presents the whole results obtained from the proposed framework.

IV. The Experimental Results

A. Data Set DCASE2016 for Audio Tagging

The study has been applied to the CHIME-HOME dataset of the DCASE 2016 For challenging audio tagging. The audio recordings were created in a domestic environment [39]. The audio data is presented as 4-second chunks at a sampling rate of 16 kHz. There are 7 labels that appear in audio segments as shown in Table II, Besides, Sounds issued from outside the house.

Alternate input features for audio tagging include MFCC and SSFT. Each audio chunk was preprocessed by segmenting it with a 20 ms sliding window and a 10 ms hop size, then converting it to 13-dimension MFCCs and 13-dimension stft with 320 window length and 160 hop size.

TABLE II. Labels of Audio Data Set

Label/Audio events	Event Description
b	Broadband noise
c	Child speech
f	Adult female speech
m	Adult male speech
o	Other identifiable sounds
p	Percussive sound events
v	Video game / TV

B. Confusion Matrix

A confusion matrix is utilized to evaluate a classifying algorithm's performance. The confusion matrix is a table that summarizes the number of accurate and inaccurate predictions that are created by the classifier (or classification model) for binary classification tasks. A confusion matrix is a N × N matrix that is utilized to assess the efficiency of a classification model, in which N as the number of target classes.

TP: True Positive: The actual value was positive and the model anticipated a positive value. FP: False Positive: Your prediction is positive, and it is false. FN: False Negative: Your prediction is negative, and result it is also false. TN: True Negative: The actual value was negative and the model predicted a negative value. Table III shows evaluation metrics that are driven from the confusion matrix.

A model with a lower EER is considered more accurate, whereas a model with a higher Accuracy Coefficient (ACC) is considered to be superior. Precision (PREC), recall (REC), and F-score are other metrics used to assess the performance of models [40].

TABLE III. The Main Evaluation Metrics

Measure	Formula
ACC	(TP+TN)/ (TP+TN+FN+FP)
ERR	(FP+FN)/ (TP+TN+FN+FP)
REC	TP / (TP+FN)
PREC	TP / (TP+FP)
F(score)	2×PREC×REC/ (PREC+REC)

C. Experiment Result Number One: Test DCASE 2016 Accuracy

This experiment is designed to evaluate the proposed framework's accuracy and loss. The data set DCASE2016 Task4 was used [41, 42]. The MFCCs and STFT features were employed. These features were trained using a deep neural network with two hidden layers: an input layer and an output layer. Adamax is used as an optimizer, and the cost function is binary cross entropy. After the time period specified, the training process comes to an end (100 epoch).

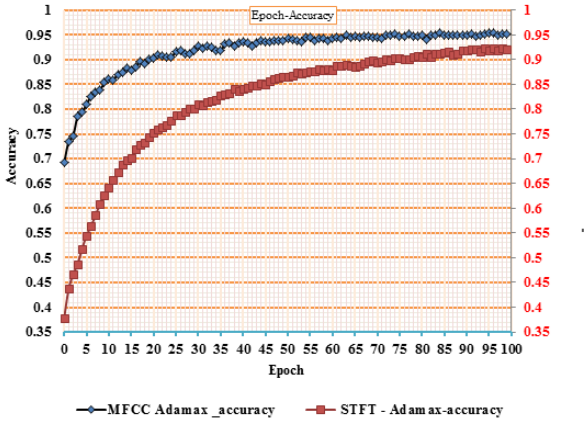


Figure 7. MFCCs Adamax accuracy and STFT Adamax accuracy by epoch.

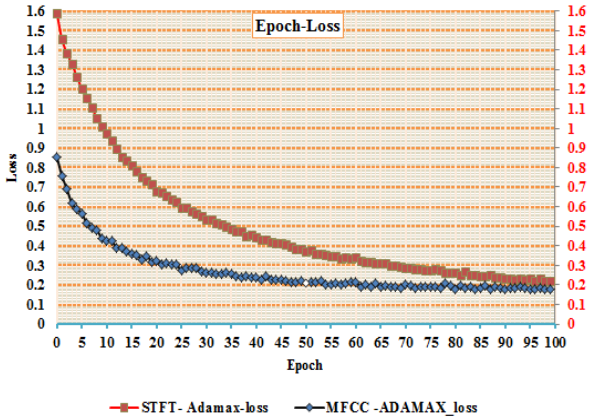


Figure 8. MFCCs Adamax loss and STFT Adamax loss by epoch.

As shown in Fig. 7, the accuracy of the feature MFCC with Adamax optimizer is better than that of the STFT with Adamax optimizer. At epoch 100, the accuracy of the MFCC-Adamax optimizer is 95% and the accuracy of the STFT Adamax optimizer is 93%. In Fig. 8, the loss of the Adamax optimizer is better than the Adam optimizer. At epoch 100, the loss of the Adamax optimizer is 0.17%. The loss of the STFT Adamax optimizer is 0.22% (see Table IV).

TABLE IV. THE LABELS OF AUDIO DATA SET

Label	Description
c	Child speech
m	Adult male speech
f	Adult female speech
v	Video game /TV
p	Percussive sounds, e.g. crash, bang, knock, footsteps
b	Broadband noise, e.g. household appliances
o	Other identifiable sounds
s	Silence / background noise only
u	Flag chunk (unidentifiable sounds, not sure how to label)

D. Experiment Result Number Two: Test Accuracy

This experiment elucidates the ACC of the proposed framework for seven tags, which measures the number of correct predictions for the framework. The results in Table V show that ACC in MFCC with Adamax optimizer outperforms ACC in STFT with Adamax optimizer. In the development set, the average accuracy increased from 0.991 to 0.994, and in the evaluation set, it increased from 0.983 to 0.966.

E. Experiment Result Number Three: Test Equal Error Rate

1) Evaluate the proposed framework

The Equal Error Rate (ERR) is tested in this experiment. Table VI displays the outcomes. For seven tags, EER from the proposed framework shows that EER in the MFCC Feature with Adamax optimizer is better than in the STFT with Adamax optimizer [43]. In the development set, the average EER decreased from 0.027 to 0.007, and in the evaluation set, it decreased from 0.023 to 0.015.

2) Overall evaluations

Table VII compares EER on seven labels between [24-28], and the proposed framework. where the proposed frame work is superior, with EER decreasing from 0.209 in [24] to 0.015 in the proposed MFCCs feature frame work.

F. Experiment Result Number Four: Test the Precision, Recall and F-Score

This experiment clarifies the tests of precision, recall, and f-score of the proposed framework. The table summarizes the performance of DCASE2016 Task4 for seven tags. As shown in Table VII, the MFCC Adamax optimizer is better than the STFT Adamax optimizer, where precision increased from 88% to 95.2%, recall increased from 88.3% to 95.2%, and F-score increased from 85.3% to 95.1%.

V. THE CHALLENGES AND LIMITATIONS OF DSCASE 2016 DATA SET

The challenges that have been met are the limitations that must be followed when dealing with data in order to improve the result and obtain a lower error rate and reduce noise.

Specified the labels used in the study: In a DSCASE challenge, for example, the audio has nine labels, as shown in Table VIII, but the maximum number of labels allowed is seven. An author can assign any subset of labels to an audio clip. They can only be set separately, with the exception of the labels S and U.

Optimizing the feature: Because each audio signal contains many features, we must select the feature that best fits the problem we want to solve. Algorithms for audio signal processing analyze signals, extract features, and detect the presence of any pattern in the signal.

The audio files for the challenge have limitations that make tagging difficult. The first source of noise is environmental noise, which occurs because the recordings are made in real-world settings. As a result, for audio tagging, a strong environmental noise resistance model is required. The second limitation is the presence of multiple sound sources in a single recording. As a result, the classification model must be capable of modelling and recognizing multiple sound sources simultaneously.

So, in our future work, we must select the optimal feature that is related to the problem and will solve it, such as the Mel filter bank, as well as the best deep learning model for a classifier, such as the denoising autoencoder, and another optimizer, such as AdamW, to improve efficiency, reduce resource consumption, and reduce time (see Table IX).

TABLE V. ACCURACY RESULTS OBTAINED FROM THE PROPOSED FRAME WORK

Tags	b	c	f	m	o	p	v	Average
Development Set								
MFCC-DNN(Adamax-opt)	0.999	0.990	0.977	0.988	0.993	0.997	0.994	0.994
STFT-DNN(Adamax-opt)	0.998	0.981	0.995	0.996	0.990	0.991	0.989	0.991
Evaluation Set								
MFCC-DNN(Adamax-opt)	0.998	0.962	0.991	0.993	0.972	0.989	0.977	0.983
STFT-DNN(Adamax-opt)	0.994	0.925	0.981	0.984	0.961	0.964	0.956	0.966

TABLE VI. EER COMPARISONS BETWEEN THE RESULTS OBTAINED FROM THE PROPOSED FRAMEWORK

Tags	b	c	f	m	o	p	v	Average
Development Set								
MFCC-DNN (Adamax-opt)	0.004	0.019	0.003	0.002	0.012	0.006	0.008	0.007
STFT-DNN (Adamax-opt)	0.005	0.080	0.010	0.008	0.030	0.020	0.040	0.027
Evaluation Set								
MFCC-DNN(Adamax-opt)	0.001	0.037	0.008	0.006	0.027	0.010	0.022	0.015
STFT-DNN(Adamax-opt)	0.003	0.061	0.012	0.009	0.023	0.018	0.039	0.023

TABLE VII. PRECISION, RECALL, AND COMPARISONS BETWEEN THE RESULTS OBTAINED FROM THE PROPOSED FRAMEWORK

	Evaluation set	b	c	f	m	o	p	v	Average
Precision	STFT-DNN(Adamax-opt)	0.983	0.804	0.917	0.908	0.848	0.869	0.836	0.880
	MFCC-DNN(Adamax-opt)	0.987	0.946	0.964	0.967	0.895	0.974	0.935	0.952
Recall	STFT-DNN(Adamax-opt)	0.980	0.629	0.941	0.994	0.891	0.894	0.855	0.883
	MFCC-DNN(Adamax-opt)	1.000	0.806	0.992	1.000	0.961	0.959	0.948	0.952
F-score	STFT-DNN(Adamax-opt)	0.981	0.706	0.894	0.935	0.819	0.818	0.819	0.853
	MFCC-DNN(Adamax-opt)	0.994	0.871	0.977	0.983	0.927	0.967	0.942	0.951

TABLE VIII. SUMMARY AND COMPARISON OF PREVIOUS STUDIES AND THE PROPOSED FRAMEWORK

Ref	**Year**	**System Characteristics**			**Equal Error Rate (Average)**		**Accuracy**
		Features	**Classifier**	**Optimizer**	**(evaluation dataset)**	**(development dataset)**	
23	2016	CQT Features	CNN	SGD	0.178	0.166	
24	2016	MFCCs	GMM		0.209	0.213	72.5%
25	2016	STFT	CNN	Adam	00.210	0.174	84.50%
26	2017	MFCCs	DNN	SGD	0.168	0.151	-
26	2017	MFBs	DNN	SGD	0.157	0.135	-
26	2017	MFBs	DAE	SGD	0.148	0.126	-
27	2016	MFBs	DNN	SGD	-	**0.209**	-
28	2016	MFCCs	RNN	ADADELTA	0.20		-
The Proposed framework	-	STFT	DNN	Adamax	0.023	0.027	93%
	-	MFCCs	DNN	Adamax	0.015	0.007	95%

TABLE IX. SUMMARY AND COMPARISON OF PREVIOUS STUDIES AND THE PROPOSED FRAMEWORK ON SEVEN LABELS ON THE EVALUATION SET

Ref.	**[24]**	**[25]**	**[26]**	**[27]**	**[28]**	**[29]**	**Proposed Framework STFT**	**Proposed Framework MFCC**
Broadband noise	0.117	0.18	0.014	0.039	0.11	0.150	0.003	0.001
Child speech	0.191	0.20	0.210	0.195	0.21	0.145	0.061	0.037
Adult female speech	0.314	0.23	0.207	0.229	0.26	0.143	0.012	0.008
Adult male speech	0.326	0.06	0.149	0.280	0.24	0.031	0.009	0.006
Other identifiable sounds	0.249	0.19	0.256	0.272	0.29	0.0135	0.023	0.027
Percussive sound events	0.212	0.11	0.175	0.221	0.23	0.013	0.018	0.010
TV sound	0.056	0.24	0.022	0.090	0.06	0.248	0.039	0.022
Average	0.209	0.17	0.148	0.189	0.20	0.123	0.023	**0.015**

VI. DISCUSSION

It is found from the Overall evaluations the following: Table VI compares EER on seven labels between [24–29] and the proposed framework. The results obtained showed that the proposed framework is superior, with EER decreasing from 0.209 in [24] to 0.015 in the proposed MFCCs feature framework.

Table V demonstrates the summary and comparison of previous studies and the proposed framework. The proposed framework processes input signals using MFCCs and STFT features to extract features or characteristics from the audio signal, and these features are then entered into the deep neural network with the Adamax optimizer. But in In Lidy and Schindler [22], a CQT feature was used and then input to CNN with the SGD optimize.In Lars and Phan *et al.* [25], a short-time Fourier transform feature was used and then fed into a convolutional neural network with masked global pooling with Adam optimizer.In Xu and Huang [26], Mfccs and mfbs were used, and then features were added to DNN and MFBs were input to the DAE. In Kong *et al.* [27], a mel filter bank feature was used and then input to DNN with the SGD optimize. In Vu and Wang [28], the MFCCs feature was used and then input to RNN with the ADADELTA optimize. Audio signal processing is very critical and challenging topic that has a great impact on different real life applications. For example, health-related activity monitoring, robotic systems, virtual assistants and Metaverse.

VII. CONCLUSIONS

In this study, we have examined the modelling and acoustic feature learning problems in audio tagging. We introduce the proposed Optimized Deep Neural Networks Audio Tagging Framework for Virtual Business Assistant trained on the DCASE 2016 data set for audio pattern recognition. MFCCs and STFT were used to extract features from the audio signal. It has been proven that MFCC feature and DNN can be used effectively for automatic labelling and classification of audio. A dropout was also used to avoid the neural network's over-fitting. Adamax is used as an optimization technique. In future work, this study will be implemented to the DAE. It might result in more high-level features being extracted for the audio tagging challenge. To more to evaluate the proposed

model, bigger datasets such as YouTube-8M dataset [44] would be regarded.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

F.E. participates in the analysis of theoretical results and in developing modeling and application programs. N.B participates in the analysis of theoretical results, preparing the theoretical model and the work of the solution program. S.E and A.E. propose the research topic and participate in the analysis of theoretical results, writing and formatting of paper, developing modeling and application programs. All authors had approved the final version.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 206–219, May 2019.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, pp. 4873–4876.

[3] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE International Conference on Multimedia and Expo (ICME*), vol. 1, pp. 452–455, 2000.

[4] S. Allegro, M. Büchler, and S. Launer, "Automatic sound classification inspired by auditory scene analysis," in *Eurospeech*, Aalborg, Denmark, September 2001.

[5] B. Picart, S. Brognaux, and S. Dupont, "Analysis and automatic recognition of human beatbox sounds: A comparative study," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2015, pp. 4255–4259.

[6] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, "Lifelogging: archival and retrieval of continuously recorded audio using wearable devices," in *Proc. IEEE International Conference on Emerging Signal Processing Applications (ESPA)*, January 2012, pp. 99–102.

[7] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, pp. 40–50, 2012.

[8] J. P. Woodard, "Modeling and classification of natural sounds by product code hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, pp. 1833–1835, 1992.

[9] D. P. W. Ellis, "Detecting alarm sounds," in *Proc. the Workshop on Consistent and Reliable Acoustic Cues (CRAC-2001)*, 2001, pp. 59–62.

[10] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, pp. 1733–1746, 2015.

[11] A. Mesaros, T. Heittola, E. Benetos, P. Foster, and M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, pp. 379–393, 2018.

[12] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017, pp. 85–92.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. DCASE2018 Workshop*, 2018, pp. 9–13.

[14] S. M. Elghamrawy and S. E. Ibrahim, "Audio signal processing and musical instrument detection using deep learning techniques," in *Proc. Japan-Egypt International Conference on Electronics, Communications and Computers (JEC-ECC)*, 2021, pp. 146–149.

[15] C. Keunwoo, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," arXiv preprint, arXiv:1606.00298, 2016.

[16] Community. [Online]. Available: http://dcase.community/challenge2016/task-audio-tagging

[17] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[18] S. Hariharan, P. Rao, and S. D. Roy, "Audio signal classification," EE Dept, IIT Bombay, pp. 1–5, 2004.

[19] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[20] N. Climent, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, no. 1–2, pp. 93–114, 2001.

[21] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1142–1158, 2009.

[22] S. Yun, S. Kim, S. Moon, J. Cho, and T. Kim, "Discriminative training of GMM parameters for audio scene classification and audio tagging," *IEEE AASP Challenge Detect. Classification Acoust. Scenes Events*, 2016.

[23] L. Thomas and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proc. the 7th Workshop on Detection and Classification of Acoustic Scenes and Events*, 2016, pp. 60–64.

[24] X. Yong, H. Qiang, W. Wenwu, J. Philip, and J. B P. D. Mark, "Fully DNN-based multi-label regression for audio tagging," arXiv preprint, arXiv:1606.07695, 2016.

[25] H. Lars, H. Phan, and A. Mertins, "Classifying variable-length audio files with all-convolutional networks and masked global pooling," arXiv preprint, arXiv:1607.02857, 2016.

[26] Y. Xu, Q. Huang, W. W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1230–1241, 2017.

[27] Q. Q. Kong, *et al*., "Deep neural network baseline for DCASE challenge 2016," in *Proc. the 7th Workshop on Detection and Classification of Acoustic Scenes and Events*, 2016.

[28] T. H. Vu and J. C. Wang, "Acoustic scene and event recognition using recurrent neural networks," in *Proc. Detection and Classification of Acoustic Scenes and Events*, 2016, pp. 1–3.

[29] Y. Xu, *et al*., "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Porc. 2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3461–3466.

[30] R. Hibare and A. Vibhute, "Feature extraction techniques in speech processing," *International Journal of Computer Applications*, vol. 107, no. 5, 2014.

[31] K. Potdar and C. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, 2017.

[32] P. J. M. Ali and R. H. Faraj, "Data normalization and standardization: A technical report," *Machine Learning Technical Reports*, no. 1, pp. 1–6, 2014.

[33] G. Hinton, L. Deng, D. Yu, G. E. Dahl, *et al*., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, November 2012.

[34] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19.

[35] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[36] H. A. Bourlard and N. Morgan, "Connectionist speech recognition: A hybrid approach," *Bourlard Springer Science & Business Media*, vol. 247, 2012.
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol.15, pp. 1929–1958, 2014.
[38] D. Yi, J. Ahn, and S. Ji, "An effective optimization method for machine learning based on ADAM," *Applied Sciences*, vol. 10, 1073, 2020.
[39] H. Christensen, N. Ma, J. P. Barker, and P. D. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," in *Proc. Eleventh Annual Conference of the International Speech Communication Association*, 2010.
[40] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS One*, vol. 10, 2015.
[41] S. M. El-Ghamrawy, A. I. El-Desouky, and M. Sherief, "Dynamic ontology mapping for communication in distributed multi-agent intelligent system," in *Porc. 2009 International Conference on Networking and Media Convergence*, 2009, pp. 103–108.
[42] S. M. El-Ghamrawy and A. I. Eldesouky, "An agent decision support module based on granular rough model," *International Journal of Information Technology & Decision Making*, vol. 11, no. 4, pp. 793–820, 2012.
[43] S. M. Elghamrawy, A. E. Hassnien, and V. Snasel, "Optimized deep learning-inspired model for the diagnosis and prediction of COVID-19," *Cmc-Computers Materials & Continua*, pp. 2353–2371, 2021
[44] A. E. H. Sami, *et al.*, "Youtube-8m: A large-scale video classification benchmark," arXiv preprint, arXiv:1609.086752, 2016.