Urban Sound Classification for Audio Analysis Using Long Short-Term Memory

Shivam Tyagi, Kanishka Aggarwal, Deepika Kumar, Shreya Garg, Neeraj
Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
shivamtyagi0918@gmail.com, kanishkaaggarwal45@gmail.com, deepika.kumar@bharatividyapeeth.edu, shreyagarg0501@gmail.com, jangraneeraj30@gmail.com

***Abstract*—**The process of audio classification involves categorizing audio signals into predefined classes based on their acoustic characteristics. Deep learning techniques have played a significant role in addressing this issue. Researchers have proposed various approaches to advance the field, including exploring different neural network architectures, incorporating auxiliary information like keywords or sentence information to guide audio classification, and implementing diverse training strategies. In this study, the researchers propose the use of a Long Short-Term Memory (LSTM) network for classifying environment sounds. The UrbanSound8K dataset's audio data files are categorized into 10 classes using the proposed LSTM model. The researchers evaluate the model using various metrics. The results show an accuracy of 0.86, precision of 0.87, recall of 0.87, support value of 1747, and an f1 score of 0.87 achieved by the proposed model. The researchers compare their methodology with state-of-the-art approaches and present the empirical evaluation alongside their findings.

***Keywords*—** *Sound Classification, Urban Sound 8K, Long Short Term Memory, Mel Spectrogram*

I. INTRODUCTION

Sound Classification is the practice of recognizing and classifying sounds according to their attributes and features. Acoustic monitoring, voice recognition, and music information retrieval are just a few of the numerous disciplines that use it as a core job. The goal of sound classification is to create algorithms that can categorize sounds. Since many years [1] ago, categorizing audio or sound has been a significant area of research, and there are numerous tried-and-true techniques with various models and features that have shown to be efficient and reliable. The technique of classifying audio signals into predetermined groups based on their acoustic characteristics is known as sound classification. The task of identifying and categorizing sounds that occur in natural or artificial surroundings, such as animal sounds, traffic noise, or human speech, is known as environmental sound classification.

Sound classification is a significant field of study with numerous [2] real-world applications. It can be used to enhance the precision of audio recognition systems and to automatically categorize songs according to their genre or mood in music information retrieval systems. It can be used in environmental monitoring systems to identify and categorize noises coming from various sources, including those made by machines, cars, and animals. The frequency range and temporal correlations are relatively little-known concepts. The three stages [3] of the sound recognition challenge are signal pre-processing, the extraction of certain features, and their classification. The input signal is split into many segments during signal pre-processing, which is used to extract associated features. Data size reduction can be achieved through the process of feature extraction. Feature extraction involves transforming complex data into compact feature vectors. By extracting meaningful features from the data, the dimensionality of the dataset can be reduced, resulting in a more concise representation of the information. However, because environmental sounds typically exhibit [4] non-stationary behavior, many linear/deterministic prediction techniques frequently fail to capture the characteristic, making performance enhancement more difficult nowadays.

A number of properties from audio signals can be extracted using signal processing techniques, and the features are then utilized to train ML models for audio classification and other applications. Spectral features [5], Temporal features [6], Mel-frequency cepstral coefficients (MFCCs) [7], Pitch and timbre [8] characteristics, Wavelet features, Harmonic-percussive separation features [9], etc. are some typical features that can be produced using signal processing techniques. In addition to features created by signal processing techniques, such as MFCCs, Discrete Wavelet Transform coefficients, and Matching Pursuit features.

The use of ML algorithms is one of the most used methods for classifying sounds. On a dataset of audio files that have been labeled with the corresponding categories, these algorithms are trained. In order to classify new sounds, the algorithms learn to identify patterns in the data that are exclusive to each category. There are several types of machine learning algorithms that can be used for sound classification, including the k-Nearest Neighbors (KNN) [10] algorithm, Support Vector Machine (SVM) [11], Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM) [12]. Deep neural networks (DNN) enable feature engineering while maintaining classification accuracy and even surpass [13] the conventional approaches, in contrast to the approaches mentioned. Convolutional neural networks (CNN) are very effective in capturing Spectro-temporal patterns from

spectrogram-like input. The other class of neural network designs used for sound categorization are CNN hybrids. CNNs, in particular, have recently gained prominence as a potent method for sound classification. Using unprocessed audio signals, CNNs may automatically extract pertinent features and learn to categorize them into several sound classes. In addition to ML and deep learning (DL) algorithms, there are also rule-based approaches to sound classification. Rule-based methods involve defining a set of rules or heuristics that can be used to classify sounds based on their properties. These approaches are often used in acoustic monitoring applications, where the goal is to detect specific sounds, such as bird calls or frog choruses. The complex and varied nature of environmental sounds, however, makes it difficult to develop an efficient CNN model for environmental sound classification.

Audio classification datasets, which consist of substantial collections of labeled audio recordings organized by their acoustic properties, are crucial for training machine learning models to identify and categorize various types of sounds. The most popular datasets for audio classification include UrbanSound8K [14], ESC-50 [15], GTZAN Genre Collection [16], Voice Commands, and AudioSet [17]. Each of these datasets has distinctive qualities of its own, with some concentrating on particular sound types, such as spoken commands, musical genres, or environmental sounds. In this study, the appropriateness of an LSTM model for classifying urban sounds using the UrbanSound8K dataset is explored and analyzed.

The following is an overview of the research contributions:

- The UrbanSound8K dataset has been used to train and evaluate an LSTM model designed for the classification of environmental sounds. The evaluation encompasses performance metrics such as accuracy, precision, support, F1 score, and recall.
- A comprehensive comparative analysis of cutting-edge methodologies recently published in the field of audio classification has been conducted. This examination reveals the strengths and weaknesses of these novel approaches, providing valuable insights for further exploration.
- Furthermore, the proposed architectural framework has been compared to state-of-the-art classification algorithms, primarily in manners of accuracy. This assessment allows for a rigorous appraisal of the model's performance and efficacy.
- The study explores the impact of diverse training strategies and the incorporation of auxiliary information on the overall performance of the LSTM model. This empirical investigation unravels the nuanced interplay between factors, enriching the understanding of the model's capabilities and limitations.
- The findings of this research contribute to the existing knowledge in the field, providing profound insights into the efficacy of the LSTM model for urban sound classification. Consequently, this work holds substantial implications for both academic researchers and practitioners seeking to advance the domain.

The paper is structured as follows: The latest research in this area is compiled in Section 2 along with a comprehensive evaluation of state-of-the-art approaches. The suggested model architecture is thoroughly illustrated in Section 3, along with every feature extraction and data pre-processing method. Following the presentation of the findings and the subsequent analysis in Section 4, Section 5 statistically examines the importance of the suggested methodology. In Section 6, along with the future scopes, conclusions have also been formed.

II. Literature Survey

There has been a tremendous amount of research in the field of audio classification using machine learning (ML). Researchers have explored the use of artificial intelligence (AI) and ML techniques to improve the accuracy of automated classification tools. In 2017, Huy Phan et al. [18] proposed a deep RNN for the purpose of classifying the environment in which the model extracts both temporal and spatial characteristics from the audio data by combining CNN and RNN layers. For the LITIS Rouen dataset, the suggested method received an F1-score of 97.7%. In [19], the authors introduced a Convolutional Recurrent Neural Network (CRNN) architecture that combines convolutional and recurrent layers to capture both local and global aspects of sound sources. An ensemble approach of CNNs [20] was employed to improve the classification performance on the hypothesis that single CNN may not be sufficient to capture all the complex features.

In the publication [21], a novel method for estimating rainfall from audio data is presented, and it is shown how well a CNN architecture does this task. This method may be used in various circumstances when estimating rainfall is crucial, such as in flood warning systems. The authors, Nithya Davis et al. [22], train numerous CNN models using various architectures and hyperparameters and assess their effectiveness using the ESC-50 dataset. The best model outperforms previous techniques for environmental sound classification, achieving an accuracy of 86.2%. For the ESC-50 dataset, a deep CNN model with a VGG-style model and a ResNet model was proposed in [23]. The models that were trained with data augmentation perform noticeably better than those that were trained on the original dataset without it.

Using the DCASE 2017 Challenge dataset, the authors [24] assessed their system and demonstrated that it outperforms conventional techniques by achieving a classification accuracy of 85%. Furthermore, they

demonstrated how the SED module can accurately detect when background noises appear in the audio segments, which is helpful for speech recognition and other audio processing tasks. The paper [25] introduced a novel approach for categorizing environmental sound using a concatenated spectrogram and a deep CNN. The findings concluded that the suggested model outperformed the alternatives and has promise for use in acoustic monitoring and soundscape analysis, among other applications. A comparison study of various semi-supervised deep learning methods for audio categorization tasks was provided by Léo Cances et al. [26]. The scientists employed two distinct datasets, UrbanSound8K and FSD50K, which both included a sizable number of audio recordings of various sound classifications, including siren, car horn, and dog barking. These datasets were subjected to the Pseudo-Labeling, Mean Teacher, and MixMatch algorithms, which are three separate semi-supervised learning techniques. According to the findings, the MixMatch algorithm is a good method for categorizing audio files and has promise for use in acoustic monitoring and sound event identification.

The study [27] suggested a method for automatically choosing features in audio categorization using spectrogram images. The suggested technique chooses a small subset of pertinent characteristics from a huge pool of features derived from the spectrogram images using a combination of two feature selection algorithms. The approach presented by Arooshi Taneja et al. [28] extracts information from cardiac sound waves and categorizes them into several groups, such as normal, murmur, and pathological. The authors of the study assessed the performance of their suggested categorization approach against those of other methods already in use using a publicly available dataset of heart sounds. Using a deep audio feature extraction strategy, a Bidirectional LSTM (BLSTM) network [29] was developed in 2018 for classifying acoustic scenes. The proposed method is assessed on the DCASE 2019 Task 1B dataset, which contains audio recordings of 10 different acoustic scenes. The findings indicated that the suggested approach, which achieves an accuracy of 83.5%, outperformed numerous state-of-the-art methods on the grounds of classification accuracy.

An approach for categorizing audio was proposed by Krishna Kumar et al. [30] that incorporates feature extraction, neural network classification, and principal component analysis (PCA). The UrbanSound8K dataset includes audio recordings of various environmental sounds and is used to assess the approach. The findings demonstrate that the suggested strategy outperforms various cutting-edge techniques, with a classification accuracy of 87.1%. An approach for categorizing audio based on fuzzy-rough nearest neighbor (FRNN) clustering is employed in [31]. To address uncertainties and inconsistencies in the data, the FRNN clustering algorithm combines fuzzy set theory and rough set theory. The ESC-50 dataset and the UrbanSound8K dataset are used to evaluate the approach. The findings demonstrate that the suggested strategy performed various novel approaches and produced excellent classification accuracies. In order to deal with uncertainties and inconsistencies in the data, the FRNN clustering method is shown to be successful, and when combined with MFCC features, it produces successful outcomes for audio classification.

To deal with the issue of different sound lengths in the dataset, the adaptive data padding approach was introduced in [32]. To make sure that all audio samples are the same length, it applies adaptive data padding to the MFCCs by inserting zeros at the beginning and end of each audio sample. The system divides the audio samples into various sound categories using a deep CNN. The padded MFCCs and their related labels are used to train the CNN. The algorithm surpasses other cutting-edge sound classification algorithms and achieves excellent classification accuracy. For identifying environmental sounds, Mohamed Bubashait et al. [33] suggested a machine learning-based method. The program effectively chooses representative samples from the dataset and increases classification accuracy by using a method known as optimum allocation sampling. The system divides the audio samples into various environmental sound categories using a deep CNN. The chosen subset of data and their related labels are used to train the CNN. The algorithm surpasses other cutting-edge sound classification algorithms and achieves excellent classification accuracy.

A machine learning-based method for identifying speech activity and classifying sound events in audio signals is the Two-Stage LSTM-Based Method for Voice Detection with Sound Classification [34]. LSTM networks are employed in the method in a two-stage process. The approach outperforms other cutting-edge techniques in vocal activity recognition and sound event categorization tasks, achieving high accuracy in both.

In recent years, there has been a lot of crucial study in the field of audio classification. For audio categorization, a number of strategies have been put forth, including deep learning-based techniques like CNNs, RNNs, and LSTM networks. The MFCCs method, which has been demonstrated to be successful in capturing the spectrum properties of audio signals, is one frequent audio feature extraction technique used in audio categorization. In order to expand the quantity and diversity of audio datasets, different data augmentation techniques, for instance, time stretching and pitch shifting, have also been applied. Many studies have also concentrated on using hierarchical methods for audio classification, such as multi-level classification and auditory event detection, as well as contextual information.

III. Proposed Methodology

The research makes substantial efforts to improve the outcomes and precision of audio classification. The authors have developed an LSTM model for the multinomial classification of audio. Before providing input

to the model, the data has undergone pre-processing. Figure 1 depicts the flow diagram for the proposed LSTM method.

A. Dataset Description

The UrbanSound8K dataset [35] is a widely utilized dataset for sound classification problems, particularly those involving urban sounds. It includes 8732 labeled sound recordings comprising 10 distinct urban sound categories, such as air conditioner, jackhammer, siren, car horn, children playing, dog barking, drilling, gunshot, engine running and street music. The dataset has a variety of sounds, which is one of its benefits. The dataset includes a wide variety of sounds that are typical of metropolitan settings, including both transient and stationary noises (such air conditioners and vehicle idle). The dataset is ideally suited for testing and training sound classification algorithms capable of handling a wide spectrum of distinct sound categories because of its diversity.

An additional benefit of the UrbanSound8K dataset provides high-quality sound recordings meticulously labeled and collected using state-of-the-art equipment. This dataset ensures accuracy and consistency, critical for precise sound classification tasks.

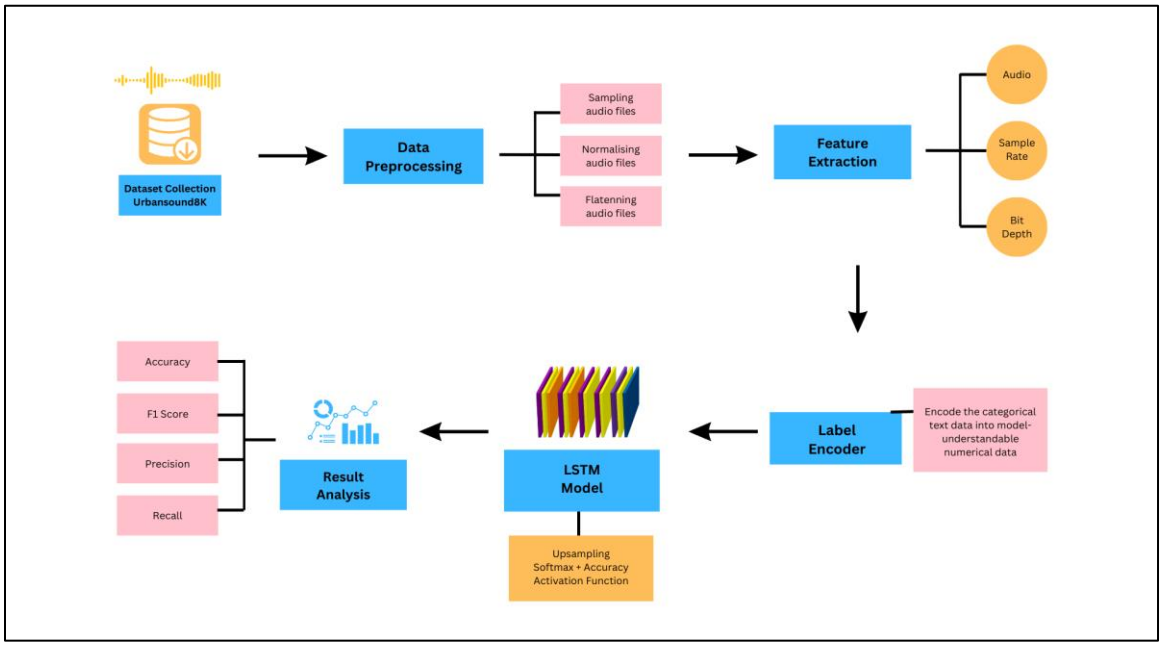

Figure 1. This figure shows the flow diagram for the proposed LSTM method.

It has been extensively employed in studies involving deep learning sound classification models, spatial analysis of urban sounds, and evaluating sound classification techniques in noisy environments. Table 1 displays the frequency distribution of audio files across various classes.

Table 1. Dataset distribution of UrbanSound8K dataset.

S. No.	Title of Audio Sample	Count of Audio Sample
1	air_conditioner	1000
2	siren	929
3	children_playing	1000
4	street_music	1000
5	drilling	1000
6	engine_idling	1000
7	car_horn	429
8	jackhammer	1000

9	gun_shot	374
10	dog_bark	1000

Several contests, notably the DCASE 2018 and 2019 challenges, have also utilized the dataset. The UrbanSound8K dataset has become a benchmark for sound classification problems as a result of these competitions, which have promoted the development of fresh and creative sound classification methods.

B. Data Pre-Processing

The audio files in the UrbanSound8K dataset were preprocessed, which involved resampling them to a constant sample rate and bit depth to standardize the audio data. The UrbanSound8K dataset required preprocessing to prepare it for sound classification tasks. This included normalizing the audio data, extracting relevant features, generating enriched versions of the data, reducing computing costs, and standardizing the audio data. Consequently, the data was preprocessed before being used in subsequent evaluations that adhered to standard practices.

The audio files were resampled to a constant sample rate of 22050 Hz to standardize the data. Then, features were extracted using the melspectrogram function, which applies a frequency-domain filter bank to audio signals, and the features from the audio files were converted into NumPy arrays. Various feature extraction techniques, such as time - domain features[36], frequency - domain features [37], and time - frequency features [38], are frequently employed in audio processing.

The audio data was converted into a spectrogram to extract the features using the melspectrogram() [39] function from the librosa library, which applies a frequency-domain filter bank to windowed audio signals. The resulting spectrogram was then converted into decibels, which computed the scaling in a numerically stable manner. The spectrogram was displayed as an image using the specshow() [40] function from librosa, with the spectrogram plotted with frequency on the y-axis and time is plotted on the x-axis. The y_axis parameter was set to 'mel' to utilize a Mel frequency scale, and the maximum frequency parameter was set to 8000 to limit the frequency range to 8 kHz. MFCCs were utilized to transform the features due to their advantageous characteristics. Subsequently, the features were normalized to enhance model convergence and accuracy. The UrbanSound8K dataset consists of 10 unique string labels; therefore, label encoding was performed to map the string labels to numerical values (0-9) using the LabelEncoder function from the Scikit-Learn library. Spectrograms depicting sets of frequencies in the audio were shown in Figure 2 (a-c).

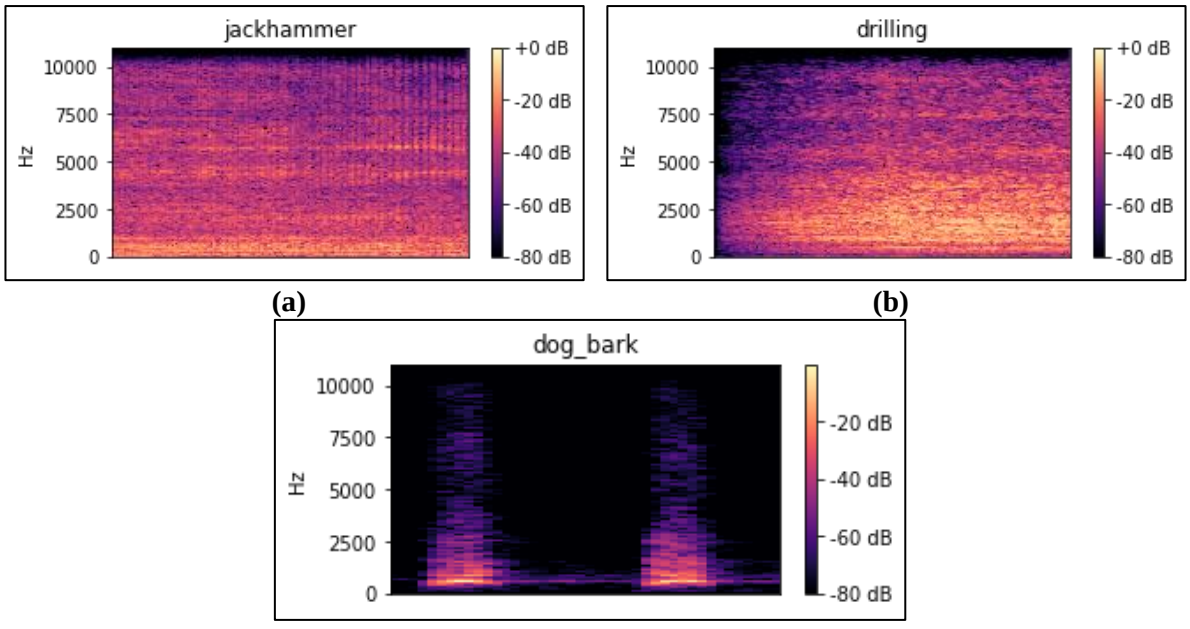

Figure 2(a) Spectrogram depicting set of frequencies in jackhammer audio. (b) Spectrogram depicting set of frequencies in drilling audio. (c) Spectrogram depicting set of frequencies in dog_bark audio.

C. Model Architecture

The UrbanSound8K dataset is a popular dataset that has been extensively utilized in studies to develop methods for the classification of urban sounds. There have been several novel approaches based on employing CNN [41], Recurrent Neural Networks (RNN) [42], SVM [43-44], Random Forest [45], K-Nearest Neighbours (KNN) [46], etc. Here, the authors have proposed employment of the LSTM model to classify audio files.

1) Long Short Term Memory (LSTM):

LSTM [47] models have been extensively used in audio classification tasks. Audio classification involves categorizing audio signals into specific classes or categories based on their acoustic characteristics. LSTM models are particularly effective in capturing temporal dependencies and long-term patterns in audio data, making them efficient for tasks such as speech recognition, music genre classification, environmental sound classification, and more. By leveraging the sequential nature of audio signals, LSTM models can learn and extract meaningful features that contribute to accurate audio classification. Their ability to handle variable-length input sequences and capture temporal dynamics makes LSTM models a popular choice in the field of audio classification.

With their recurrent structure and memory cells, LSTM models excel at capturing temporal dependencies and modelling sequential patterns in audio data making them highly suitable for tasks such as audio event detection, speech recognition, and sound classification. By processing audio signals over time and retaining long-term contextual information, LSTM models can effectively differentiate between various audio classes based on their acoustic characteristics. Their ability to learn from sequential data and adapt to different audio contexts has led to significant advancements in audio classification research, enabling more accurate and robust classification of diverse audio signal.

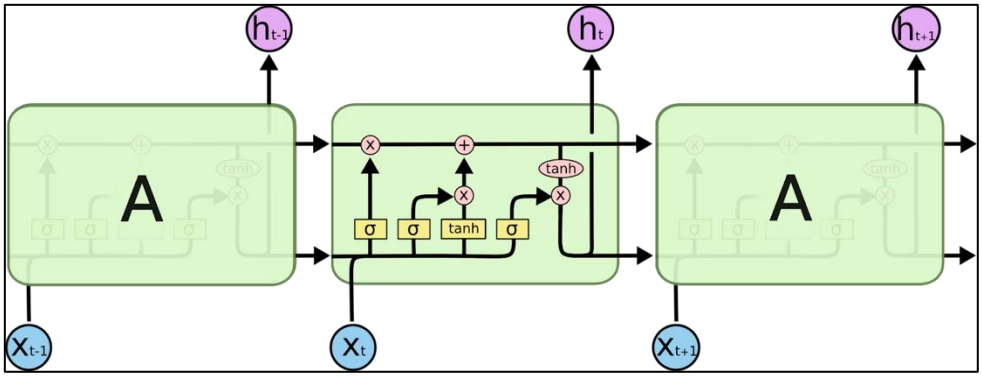

Figure 3. An overview of the LSTM architecture [48]

2) Proposed LSTM Algorithm:

The proposed methodology utilizes an LSTM model comprising two LSTM layers. These layers have sizes of 128 and 64 units, respectively. The input shape of the model corresponds to the shape of the training data, including the number of MFCC coefficients, time steps, and channels. It is important to consider the input vector's size as it affects the number of parameters in the network and the computational complexity of training and inference. Dropout regularization is employed in the LSTM layers with a rate of 0.2 to prevent overfitting and facilitate learning robust features.

The model's output layer consists of a dense layer with 10 units and utilizes a softmax function to predict class probabilities. To compile the model, the sparse_categorical_crossentropy loss function is used since the labels are integers ranging from 0 to 9. The Adam optimizer is employed. During training, the model is trained for 50 epochs using a batch size of 32. The performance of the model is evaluated using the validation data after each epoch. The architecture of this methodology is shown in Figure 4.

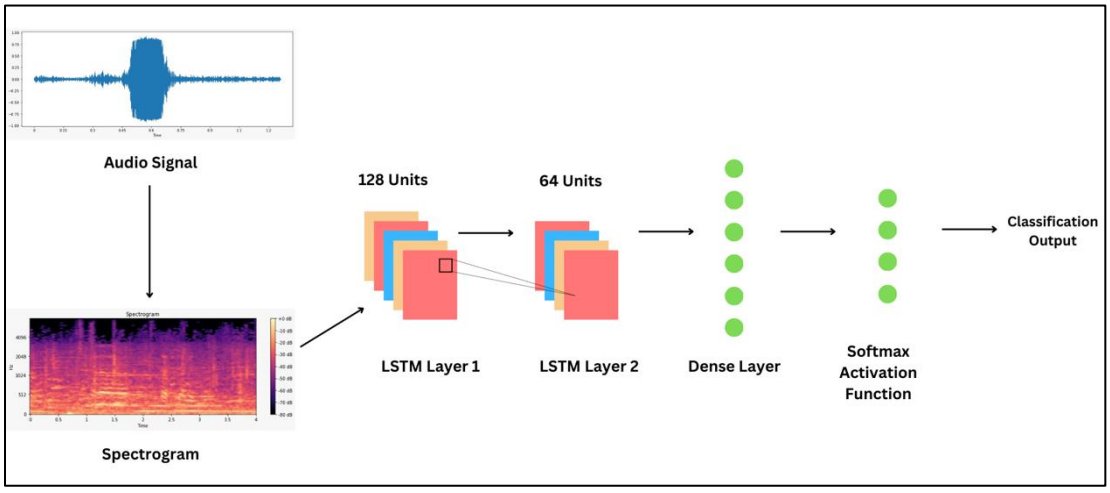

Figure 4. Architecture of proposed methodology.

IV. Results and Analysis

The LSTM algorithm was utilized in the proposed methodology. The dataset was divided into training and testing sets, with 80% and 20% of the data, respectively. To assess the performance of the algorithms, commonly used evaluation metrics such as Accuracy, Precision, Recall, Support, and F1 score were employed.

The UrbanSound8K dataset was employed to train and test the proposed architecture. It was observed that the model achieved convergence after 50 epochs. The evaluation metrics used to assess the model's performance included Accuracy, Precision, Recall, Support, and F1 score. Precision, also referred to as positive predictive value, is calculated as the ratio of true positives to the sum of false positives and true negatives. Recall, also known as sensitivity or specificity, is computed as the ratio of correctly predicted outcomes to all predictions. Accuracy represents the ratio of correct predictions to the total number of predictions made by the algorithm.

True Positives (TP) refers to cases where both the actual class and predicted class of a data point are 1. True Negatives (TN) are instances where both the actual class and predicted class of a data point are 0. False Positives (FP) occur when the actual class of a data point is 0, but the predicted class is 1. False Negatives (FN) arise when the actual class of a data point is 1, but the predicted class is 0.

1. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
2. $Precision = \frac{TP}{TP+FP}$
3. $Recall = \frac{TP}{TP+FN}$
4. $F1\ Score = \frac{2*recall*precision}{precision+recall}$

The proposed model demonstrated an accuracy of 0.86, precision of 0.87, recall of 0.87, support of 1747, and an F1 score of 0.87. A comprehensive analysis of the evaluation metrics for all classes is presented in Table 2.

Table 2. Evaluation of metrics on label classes of Urban Sound 8K Dataset

Label	Precision	Recall	F1 Score
0	.88	.92	.90
1	.89	.86	.88
2	.80	.84	.82
3	.82	.85	.83
4	.87	.84	.86
5	.90	.98	.94
6	.89	.88	.88
7	.94	.90	.92
8	.92	.88	.90
9	.81	.73	.77

Table 3 presents a comparative analysis of the results to assess the performance of the proposed LSTM model. It is evident from the analysis that the accuracy of the LSTM model surpasses all other models, indicating its superior performance. Furthermore, a confusion matrix is provided to visualize the accuracy of the model's predictions on the test set. The confusion matrix, denoted as C, represents the number of data points belonging to class i that are predicted to be in class j, with Ci,j being the corresponding value. Figure 5 illustrates the confusion matrix for reference.

Table 3. Comparative Analysis of Urban Sound Detection Model.

Model	Accuracy
LSTM	81.96%
ANN	78.24%
CNN	76.24%
Enhanced LSTM	86.7%

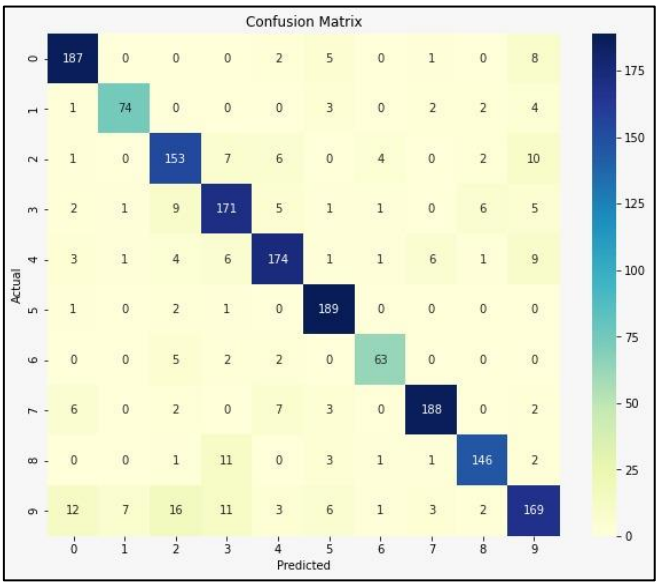

Figure 5. The figure shows the confusion matrix of the LSTM model for audio classification

The dataset was employed to train the model for a total of 50 epochs. During the evaluation process, both accuracy and loss functions were considered. The model was used to predict the desired outcome on the evaluation dataset, and the resulting predictions were compared to the expected outcomes. This comparison allowed for a real-world assessment of the model's performance. Figure 6a and 6b provide a visual comparison of the accuracies achieved by the model when applied to both the training and testing datasets.

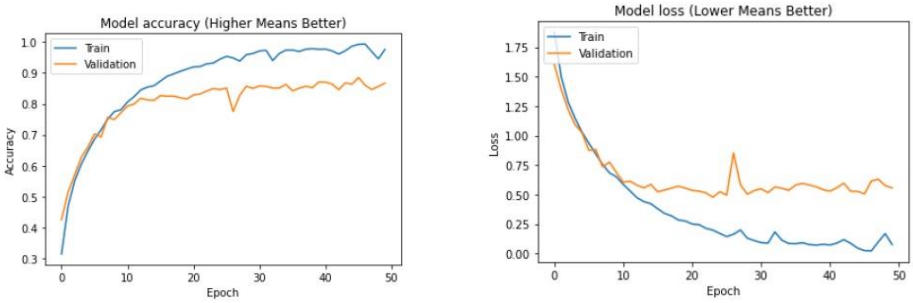

Figure 6 (a), (b). The provided image illustrates a comparison of the model's accuracies achieved when applied to both the training and testing datasets.

V. CONCLUSION

Deep learning methods, particularly LSTM networks, have garnered significant attention in recent years for their potential in audio classification. The rise in audio data volume and complexity has necessitated the need for models capable of handling sequential input and understanding long-term dependencies, which is where LSTM networks excel. The effectiveness of LSTM networks in audio categorization depends on factors such as the quality and diversity of training data, feature selection, and neural network architecture. Signal processing techniques like spectrogram analysis, MFCCs, and wavelet transforms can be employed to process the audio data and extract meaningful characteristics. Once the features are extracted, they can be inputted into an LSTM model, which can be trained to accurately classify different types of audio. LSTM networks have shown success in tasks such as speech recognition, acoustic event detection, etc. Although utilizing LSTM networks for audio classification presents challenges including the need for large amounts of data and computational power, the potential rewards are significant. Due to their ability to handle sequential data, LSTM networks are well-suited for various audio classification applications and have demonstrated excellent accuracy in classifying audio data. Further research and evaluation of the proposed model across different domains could potentially surpass existing benchmarks and enhance the current state-of-the-art in audio processing and classification.

REFERENCES

[1] Das, J. K., Ghosh, A., Pal, A. K., Dutta, S., & Chakrabarty, A. (2020, October 21). Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features. *2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS)*. https://doi.org/10.1109/icds50568.2020.9268723

[2] Kumar, R., Gupta, M., Ahmed, S., Alhumam, A., & Aggarwal, T. (2022). Intelligent Audio Signal Processing for Detecting Rainforest Species Using Deep Learning. *Intelligent Automation & Soft Computing*, *31*(2).

[3] Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. *IEEE Access*, *7*, 7717–7727. https://doi.org/10.1109/access.2018.2888882

[4] Gupta, A., Kumar, R., & Kumar, Y. (2022, December). An Automatic Speech Recognition System: A systematic review and Future directions. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 1492-1496). IEEE.

[5] Gupta, A., Kumar, R., & Kumar, Y. (2023). An automatic speech recognition system in Indian and foreign languages: A state-of-the-art review analysis. *Intelligent Decision Technologies*, (Preprint), 1-19.

[6] Ibrahim, N., Jamal, N., Sha'abani, M. N. A. H., & Mahadi, L. F. (2021, March 1). A Comparative Study of Heart Sound Signal Classification Based on Temporal, Spectral and Geometric Features. *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. https://doi.org/10.1109/iecbes48179.2021.9398810

[7] MEL-FREQUENCY CEPSTRAL COEFFICIENTS FOR SPEAKER RECOGNITION : A REVIEW. (2015, May 31). *International Journal of Advance Engineering and Research Development*, *2*(05). https://doi.org/10.21090/ijaerd.0205157

[8] Zotkin, D., Shamma, S., Ru, P., Duraiswami, R., & Davis, L. (2003). Pitch and timbre manipulations using cortical representation of sound. *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*. https://doi.org/10.1109/icme.2003.1221328

[9] de Lima Aguiar, R., e Gomes da Costa, Y. M., & Nanni, L. (2016, October). Music genre recognition using spectrograms with harmonic-percussive sound separation. *2016 35th International Conference of the Chilean Computer Science Society (SCCC)*. https://doi.org/10.1109/sccc.2016.7836027

[10] Andarabi, S., Nobakht, A., & Rajebi, S. (2020, June). The Study of Various Emotionally-sounding Classification using KNN, Bayesian, Neural Network Methods. *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. https://doi.org/10.1109/icecce49384.2020.9179451

[11] Sen, I., Saraclar, M., & Kahya, Y. P. (2015, July). A Comparison of SVM and GMM-Based Classifier Configurations for Diagnostic Classification of Pulmonary Sounds. *IEEE Transactions on Biomedical Engineering*, *62*(7), 1768–1776. https://doi.org/10.1109/tbme.2015.2403616

[12] Yamashita, M. (2021, August 23). Classification Between Normal and Abnormal Respiration Using Ergodic HMM for Intermittent Abnormal Sounds. *2021 29th European Signal Processing Conference (EUSIPCO)*. https://doi.org/10.23919/eusipco54536.2021.9616313

[13] Lezhenin, I., Bogach, N., & Pyshkin, E. (2019, September 26). Urban Sound Classification using Long Short-Term Memory Neural Network. *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. https://doi.org/10.15439/2019f185

[14] Salamon, J., Jacoby, C., & Bello, J. P. (2014, November 3). A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the 22nd ACM International Conference on Multimedia*. https://doi.org/10.1145/2647868.2655045

[15] Piczak, K. J. (2015, October 13). ESC. *Proceedings of the 23rd ACM International Conference on Multimedia*. https://doi.org/10.1145/2733373.2806390

[16] Tzanetakis, G., & Cook, P. (2002, July). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302. https://doi.org/10.1109/tsa.2002.800560

[17] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017, March). Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2017.7952261

[18] Phan, H., Koch, P., Katzberg, F., Maass, M., Mazur, R., & Mertins, A. (2017, August 20). Audio Scene Classification with Deep Recurrent Neural Networks. *Interspeech 2017*. https://doi.org/10.21437/interspeech.2017-101

[19] Sang, J., Park, S., & Lee, J. (2018, September). Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms. *2018 26th European Signal Processing Conference (EUSIPCO)*. https://doi.org/10.23919/eusipco.2018.8553247

[20] Nanni, L., Costa, Y. M. G., Aguiar, R. L., Mangolin, R. B., Brahnam, S., & Silla, C. N. (2020, May 26). Ensemble of convolutional neural networks to improve animal audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2020*(1). https://doi.org/10.1186/s13636-020-00175-3

[21] Avanzato, R., Beritelli, F., Di Franco, F., & Puglisi, V. F. (2019, September). A Convolutional Neural Networks Approach to Audio Classification for Rainfall Estimation. *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. https://doi.org/10.1109/idaacs.2019.8924399

[22] Davis, N., & Suresh, K. (2018, December). Environmental Sound Classification Using Deep Convolutional Neural Networks and Data Augmentation. *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. https://doi.org/10.1109/raics.2018.8635051

[23] Salamon, J., & Bello, J. P. (2017, March). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/lsp.2017.2657381

[24] Singh, J., & Joshi, R. (2019, October). Background Sound Classification in Speech Audio Segments. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. https://doi.org/10.1109/sped.2019.8906597

[25] Chi, Z., Li, Y., & Chen, C. (2019, October). Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. https://doi.org/10.1109/iccsnt47585.2019.8962462

[26] Cances, L., Labbé, E., & Pellegrini, T. (2022, September 19). Comparison of semi-supervised deep learning algorithms for audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2022*(1). https://doi.org/10.1186/s13636-022-00255-6

[27] Zeng, Y., Mao, H., Peng, D., & Yi, Z. (2017, December 26). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, *78*(3), 3705–3722. https://doi.org/10.1007/s11042-017-5539-3

[28] Kumar, D., & Batra, U. (2021). Classification of Invasive Ductal Carcinoma from histopathology breast cancer images using Stacked Generalized Ensemble. *Journal of Intelligent & Fuzzy Systems*, *40*(3), 4919-4934.

[29] Li, Y., Li, X., Zhang, Y., Wang, W., Liu, M., & Feng, X. (2018, July). Acoustic Scene Classification Using Deep Audio Feature and BLSTM Network. *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. https://doi.org/10.1109/icalip.2018.8455765

[30] Kumar, K., & Chaturvedi, K. (2020, February). An Audio Classification Approach using Feature extraction neural network classification Approach. *2nd International Conference on Data, Engineering and Applications (IDEA)*. https://doi.org/10.1109/idea49133.2020.9170702

[31] Wei Yang, Xiaoqing Yu, Jijun Deng, Xueqian Pan, & Yunhui Wang. (2011). Audio classification based on fuzzy- rough nearest neighbor clustering. *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2011)*. https://doi.org/10.1049/cp.2011.0901

[32] Qin, W., & Yin, B. (2022, January). Environmental Sound Classification Algorithm Based on Adaptive Data Padding. *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*. https://doi.org/10.1109/scset55041.2022.00028

[33] Bubashait, M., & Hewahi, N. (2021, September 29). Urban Sound Classification Using DNN, CNN & LSTM a Comparative Approach. *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. https://doi.org/10.1109/3ict53449.2021.9581339

[34] Feng, Y., Liu, Z. J., Ling, Y., & Ferry, B. (2022, January 7). A Two-Stage LSTM Based Approach for Voice Activity Detection with Sound Event Classification. *2022 IEEE International Conference on Consumer Electronics (ICCE)*. https://doi.org/10.1109/icce53296.2022.9730179

[35] Salamon, J., Jacoby, C., & Bello, J. P. (2014, November 3). *UrbanSound8K*. Zenodo. https://doi.org/10.5281/zenodo.1203745

[36] Hertel, L., Phan, H., & Mertins, A. (2016, July). Comparing time and frequency domain for audio event recognition using deep learning. *2016 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn.2016.7727635

[37] Alene, Y. D., & Beyene, A. M. (2020, July). Frequency-domain Features for Environmental Accident Warning Recognition. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. https://doi.org/10.1109/worlds450073.2020.9210357

[38] Chu, S., Narayanan, S., & Kuo, C. C. J. (2009, August). Environmental Sound Recognition With Time–Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1142–1158. https://doi.org/10.1109/tasl.2009.2017438

[39] *librosa.feature.melspectrogram — librosa 0.10.1dev documentation*. (n.d.). Librosa.Feature.Melspectrogram &Mdash; Librosa 0.10.1dev Documentation. https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html

[40] *librosa.display.specshow — librosa 0.10.1dev documentation*. (n.d.). Librosa.Display.Specshow &Mdash; Librosa 0.10.1dev Documentation. https://librosa.org/doc/main/generated/librosa.display.specshow.html

[41] Ozer, I., Ozer, Z., & Findik, O. (2018, January). Noise robust sound event classification with convolutional neural network. *Neurocomputing*, *272*, 505–512. https://doi.org/10.1016/j.neucom.2017.07.021

[42] Mkrtchian, G., & Furletov, Y. (2022, June 29). Classification of Environmental Sounds Using Neural Networks. *2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)*. https://doi.org/10.1109/synchroinfo55067.2022.9840922

[43] Dhanalakshmi, P., Palanivel, S., & Ramalingam, V. (2009, April). Classification of audio signals using SVM and RBFNN. *Expert Systems With Applications*, *36*(3), 6069–6075. https://doi.org/10.1016/j.eswa.2008.06.126

[44] Qin, Y. P., Qin, P. D., Wang, Y., & Lun, S. X. (2013, August). A New Optimal Binary Tree SVM Multi-Class Classification Algorithm. *Applied Mechanics and Materials*, *373–375*, 1085–1088. https://doi.org/10.4028/www.scientific.net/amm.373-375.1085

[45] Ansari, M. R., Tumpa, S. A., Raya, J. A. F., & Murshed, M. N. (2021, September 14). Comparison between Support Vector Machine and Random Forest for Audio Classification. *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*. https://doi.org/10.1109/icecit54077.2021.9641152

[46] Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, & Cheng-Shu Hsu. (2006). Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. https://doi.org/10.1109/ijcnn.2006.246644

[47] Hochreiter, S., & Schmidhuber, J. (1997, November 1). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[48] Understanding LSTM Networks -- colah's blog. (n.d.). Understanding LSTM Networks -- Colah's Blog. http://colah.github.io/posts/2015-08-Understanding-LSTMs/