

Comparative Study of Optimized ANN and SVM Models for Environmental Sound Classification with Practical Web-Based Deployment

Mayishat Altab Mridu¹, Md Masum Kawsar²

^{1, 2} Department of Computer Science & Engineering, Daffodil International University, Bangladesh.

To Cite this Article: Mayishat Altab Mridu¹, Md Masum Kawsar², “Comparative Study of Optimized ANN and SVM Models for Environmental Sound Classification with Practical Web-Based Deployment”, Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 904-910.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Environmental Sound Classification (ESC) has already become an important research area in artificial intelligence because of its broad scale of applications in Internet of Things (IoT) systems, intelligent surveillance, smart cities, industrial monitoring and healthcare. This study is performing, a comparative analysis between a Support Vector Machine (SVM) and an Artificial Neural Network(ANN) for multiclass classification of environmental sounds using the UrbanSound8K benchmark dataset. A standardized preprocessing pipeline was developed whereby each audio recording was processed using the librosa library to extract 40 Mel-Frequency Cepstral Coefficients (MFCCs) [4], [15], [16]. The mean and standard deviation of the extracted 40 MFCC coefficients was computed to generate an 80-dimensional acoustic feature vector, which was normalized using Standard Scaler prior to training both models [5]. The ANN architecture was optimized using KerasTuner with the random search algorithm [14] and the SVM employed Radial Basis Function (RBF) kernel [12]. The accuracy, precision, recall, F1- score, confusion matrix, and classification report were utilized to evaluate performance. The overall classification accuracy for the optimized ANN was 94.81%, as compared to 94.66% for the SVM. To demonstrate practical applicability, the optimized ANN was deployed as a Flask-based web application named UrbanSoundAI where users can upload WAV audio recordings and obtain real-time environmental sound predictions with waveforms, Mel spectrogram, and MFCC heatmap visualizations. The results confirm an accurate, repeatable, and practical deployability of the proposed method for environmental sound classification. Moreover, it maintains efficient inference performance for real-world acoustic monitoring applications.

Keywords: Environmental Sound Classification (ESC), Artificial Intelligence, Support Vector Machine (SVM), Artificial Neural Networks(ANN), UrbanSound8K, Mel-Frequency Cepstral Coefficients (MFCC), Radial Basis Function (RBF) kernel, Web-Based Deployment, Real-time Acoustic Monitoring.

I. INTRODUCTION

The development of environmental sound classification (ESC) and its applications in intelligent monitoring, smart cities, healthcare, industrial control and automation, surveillance, and IoT systems have made it an important research area in artificial intelligence [1], [17]. In contrast to speech recognition, which is used for a small number of sounds, ESC is a pattern recognition system that is capable of recognizing naturally occurring sounds, including sirens, dog barks, drilling, gunshots, engine idling, and street music, thus allowing intelligent systems to listen to and interpret the environment around them and to assist in making automatic decisions [1]. In the last few years, the field of environmental sound classification has benefited significantly from recent developments in machine learning and deep learning [8], [9]. Support Vector Machines (SVMs) rely on handcrafted acoustic features as a traditional machine learning technique. Conversely, deep learning models learn hierarchical feature representations automatically from the extracted MFCC features, enabling effective nonlinear classification [8][9][12]. Artificial Neural Networks(ANNs) have received much interest due to their ability to achieve an optimum trade-off between performance and computation in classification [8], [9] and under standardized experimental conditions, the UrbanSound8K dataset has proven to be one of the most popular benchmark datasets for assessing environmental sound classification models [1]. In this study, an optimized ANN and an SVM are comparatively evaluated in terms of the UrbanSound8K dataset. A standardized preprocessing pipeline was created to extract 40 Mel-Frequency Cepstral Coefficients (MFCCs) from individual audio recordings [4], [15], [16]. The mean and standard deviation of the extracted coefficients were computed to form an 80-dimensional acoustic feature vector, and both models were trained using StandardScaler to normalize the feature vector [5]. The ANN architecture was fine-tuned using KerasTuner with the Random Search algorithm, and the SVM utilized the Radial Basis Function (RBF) kernel [12], [14]. Proof of concept, the best performing ANN was then finally deployed as a web app (UrbanSoundAI) that includes the visualization of a waveform, Mel spectrogram, and MFCCs heatmap, allowing for real-time classification of sounds from the environment. This study introduces the preprocessing framework based on MFCC, a fair comparative work of ANN and SVM under the same experimental settings, hyperparameter tuning of the ANN, web deployment, and extensive performance testing using various classification metrics.

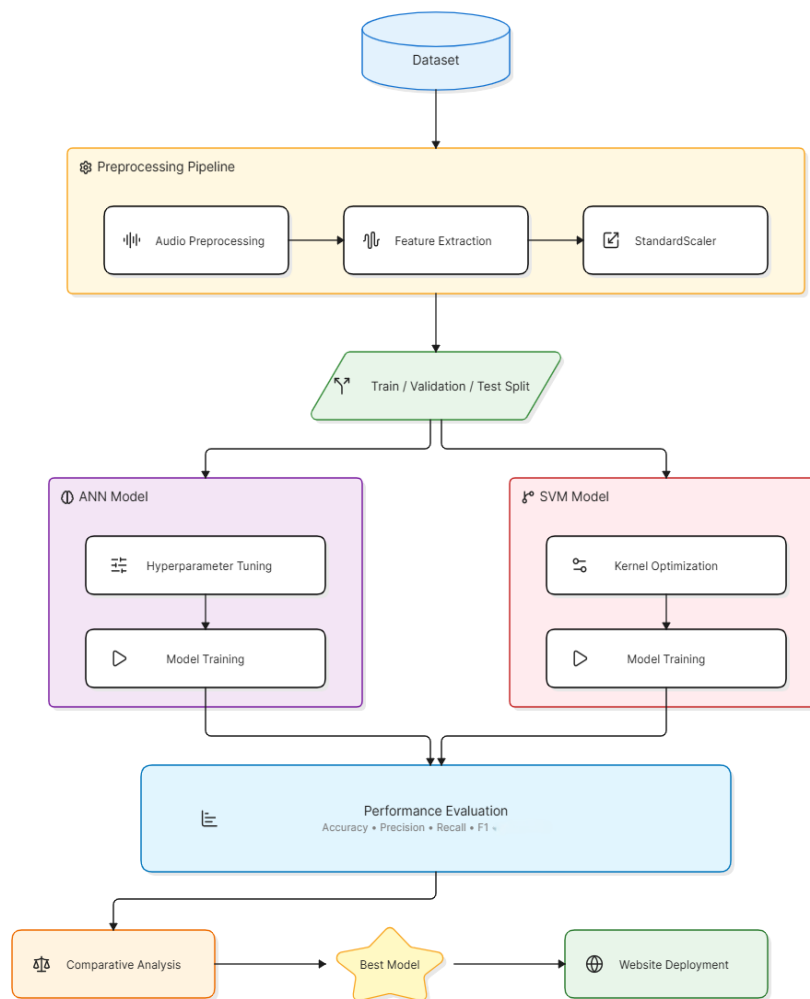


Figure 1. Overall Proposed Environmental Sound Classification Framework

II.RELATED WORK

Environmental Sound Classification (ESC) has gained much interest because of its applications, particularly in the fields of intelligent monitoring, healthcare, industrial automation, and smart city systems. The UrbanSound8K dataset has been known as one of the most popular benchmark datasets for testing environmental sound classification algorithms [1]. Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most suitable acoustic features for environmental sound analysis due to the compact and discriminative nature of the representation of audio signals [15], [16]. Davis and Mermelstein have shown their effectiveness, and Ellis has elaborated on their discussion [15], [16]. Typically, Support Vector Machines (SVMs) with handcrafted features are often used in traditional ESC systems, utilizing nonlinear kernel functions to classify effectively [11], [12]. Deep learning approaches, in recent times, have been found to perform better. Piczak et al. and Tokozume and Harada demonstrated that convolutional neural networks (CNNs) can be trained to learn the discriminative representations of environmental sounds [2], [3]. Goodfellow et al. and Géron provide the theoretical principles of these techniques[8], [9]. Despite these developments, there are not many studies that comparatively evaluate optimized SVMs and ANNs with the same preprocessing pipeline and evaluation protocol. Furthermore, there is often a lack of practical implementation. To overcome these drawbacks, this study uses a standardized MFCC-based preprocessing framework, compares the performance of ANN and SVM under the same experimental conditions, and implements the best model as an environmental sound classification web application.

III.METHODOLOGY

Overall Framework: This proposed environmental sound classification framework has six steps:

1. Dataset acquisition
2. Audio preprocessing
3. Acoustic feature extraction
4. Feature normalization
5. Model development
6. Web-based deployment

Both Artificial Neural Networks(ANN) and Support Vector Machines (SVM) have the same preprocessing steps to provide them a fair comparative evaluation.

Dataset: Experiments were performed with the UrbanSound8K benchmark dataset [1], which consists of 8,732 labeled environmental sound recordings from 10 urban sound categories. All recordings are saved as WAV files, and the maximum length is about four seconds. Although the UrbanSound8K dataset provides predefined folds for standardized evaluation, this study employed a stratified random split with a 70% training data, 15% validation data, and 15% testing data to ensure balanced class distributions across all subsets and provide a fair comparative evaluation of the ANN and SVM models.

Property	Value
Dataset	UrbanSound8K
Audio Samples	8,732
Categories	10
Audio Format	WAV
Maximum Duration	4 Seconds

Table 1. UrbanSound8K Dataset Summary

Audio Preprocessing and Feature Extraction: A standard preprocessing system was designed based on the Librosa library [4]. Each audio recording was loaded, and Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to yield a compact representation of the spectral characteristics of the environmental sounds [15], [16]. Forty MFCC coefficients were extracted for all recordings. The average and standard deviation of each coefficient were computed over all frames to obtain a fixed-length representation for machine learning. This procedure resulted in an 80-dimensional acoustic feature vector, including 40 means and 40 standard deviations. Audio data augmentation methods including additive pitch shifting, time stretching, Gaussian noise, and gain adjustment were applied only to the training data before feature extraction to improve both the generalization capability and robustness of the proposed framework. To maintain a fair and unbiased evaluation of model performance, the validation and testing datasets remained unchanged.

Artificial Neural Network(ANN): An Artificial Neural Network(ANN) was designed using the TensorFlow-Keras framework [6], [7]. The network takes an 80-dimensional feature vector as input and classifies the data into multiple classes using a Softmax output layer. The ANN architecture was optimized using KerasTuner with the Random Search algorithm to enhance classification accuracy [14]. The final architecture was composed by two hidden layers with 480 and 384 neurons, respectively. The hidden layers are followed by the batch normalization, ReLU activation function for generalization, and dropout of 0.2 to prevent overfitting. The model was trained with categorical cross-entropy loss and the Adam optimizer [13]. Early stopping and ReduceLROnPlateau were used to prevent unnecessary training and adjust the learning rate, while ModelCheckpoint was used to save the best-performing model.

Hyperparameter	Value
Hidden Layers	2
Layer 1 Neurons	480
Layer 2 Neurons	384
Activation	ReLU
Dropout	0.20
Optimizer	Adam
Learning Rate	0.0001

Table 2. Optimized ANN Hyperparameters

Support Vector Machine(SVM): The machine learning baseline model is implemented as a Support Vector Machine (SVM) for classification. For the classifier, hyperparameter optimization was performed using GridSearchCV with cross-validation. The optimal Support Vector Machine employed a Radial Basis Function (RBF) kernel [12] with a regularization parameter of $C = 100$ and $\gamma = \text{scale}$. The same normalized feature vectors as in the ANN were fed into the SVM before training, resulting in the same experimental conditions. The dataset has 10 classes for sounds; therefore, the SVM used the One-vs-One multiclass approach from Scikit-learn [5]. The setup allowed for an objective evaluation of the optimized ANN and a traditional machine learning-based classifier.

IV.EXPERIMENTAL RESULTS AND DISCUSSION

Experimental Setup: The suggested framework was implemented in Python using TensorFlow-Keras for ANN development, implementation of SVM using Scikit-learn, feature normalization, and performance evaluation, Audio Preprocessing and MFCC feature extraction using Librosa [4-7]. To ensure a fair comparison between the two classifiers, an identical preprocessing pipeline and the same training, validation, and testing datasets were used throughout all experiments.

Performance of the Optimized ANN: After optimizing the hyperparameters on the UrbanSound8K dataset, the artificial neural network performed very well in classification. Training the optimized model was shown to be very efficient, with consistent validation scores. The three techniques of batch normalization, dropout, and early stopping were effective in reducing overfitting and enhancing the generalization ability of the network. The final trained ANN was able to achieve an overall test accuracy of 94.81%. The model provided accurate results and yielded similar precision, recall, and F1 score results across all environmental sound classes, demonstrating that the model can reliably recognize environmental sounds.

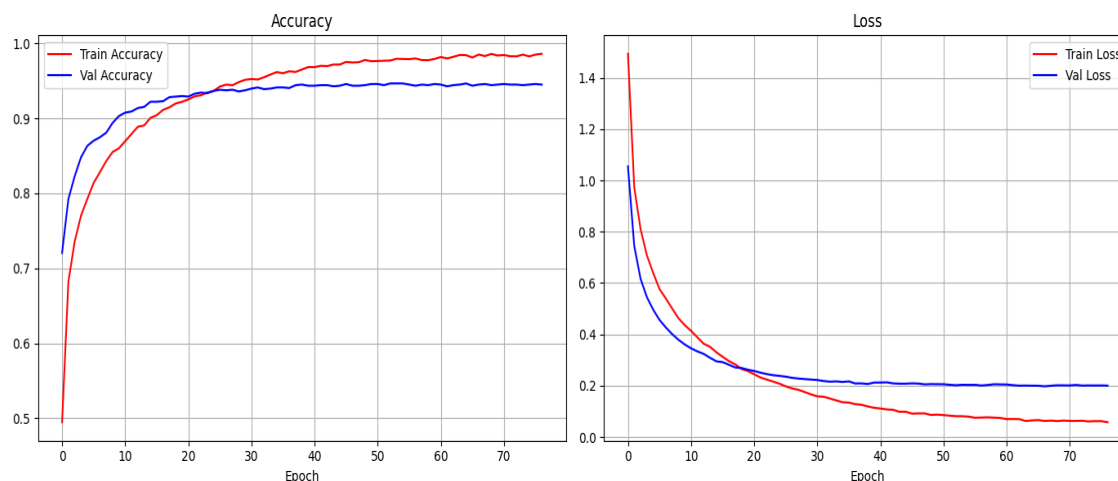


Figure 2. Training and Validation Performance of the Optimized ANN

Performance of the Support Vector Machine(SVM): The SVM with the Radial Basis Function kernel demonstrated good classification performance as well. Feature normalization was performed using the StandardScaler method before model training to ensure consistent feature scaling. The overall test accuracy is 94.66% and the SVM successfully learned the nonlinear decision boundaries in the 80-dimensional MFCC feature space. The results showed that the proposed feature representation based on MFCC has high class separability, and the proposed method enable effective used for environmental sound classification, even when a conventional machine learning algorithm is used. The SVM was less complex in terms of architecture than the optimized ANN, but its classification performance was still competitive.

Comparative Performance Analysis:

Metric	ANN	SVM
Test Accuracy	94.81%	94.66%
Precision (Macro)	0.9487	0.9431
Recall (Macro)	0.9467	0.9476
F1-score (macro)	0.9472	0.9451
Weighted F1-score	0.9479	0.9465
Training Time	572.00 s	177.39 s
Inference Time	0.7279 s	0.7973 s

Table 3. Performance Comparison of ANN and SVM

Table 3 shows the comparative performance of the optimized ANN and SVM models. These experimental results demonstrate that both classifiers achieved high classification accuracy and strong overall performance. The optimized ANN slightly outperformed the SVM in terms of accuracy, macro precision, macro F1-score, and weighted F1-score, while the SVM achieved a marginally higher macro recall. Although the performance difference between the two models was relatively small, the optimized ANN demonstrated better generalization capability while maintaining efficient inference performance. These results indicate that the optimized ANN is a more suitable model for environmental sound classification using the proposed MFCC-based feature representation.

Confusion Matrix Analysis: The confusion matrices show that both models were able to classify most environmental sound categories. There were more prediction errors between acoustically similar classes (e. g. , between air conditioner and engine idling) or between classes having similar spectral characteristics (e. g. , between drilling and jackhammer). Overall, the ANN produced fewer classification errors than the SVM, demonstrating its superior ability to learn complex nonlinear relationships from the MFCC feature representation.

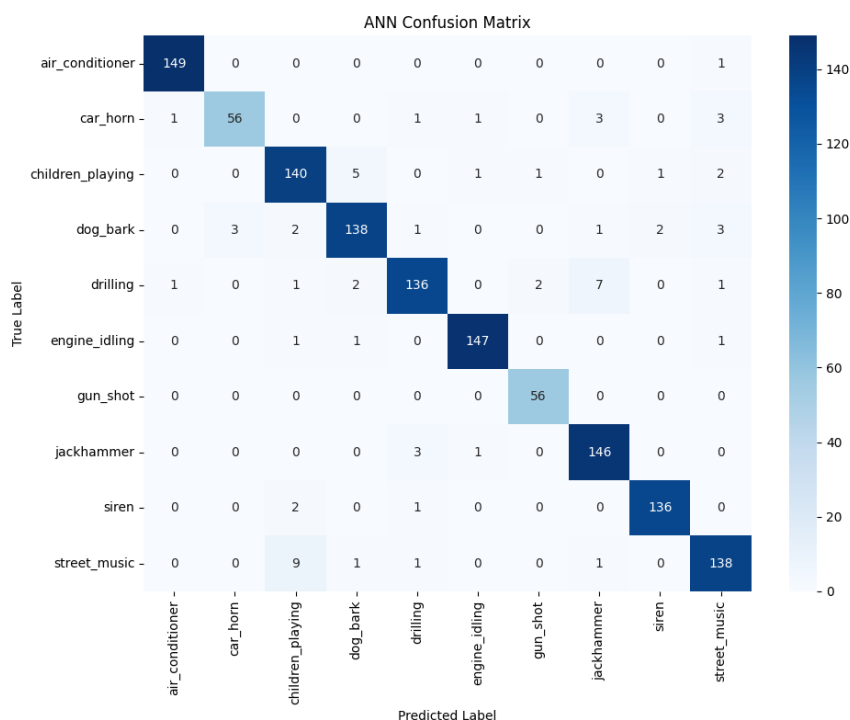


Figure 3. Confusion Matrix of the Optimized ANN

Web-Based Deployment: To illustrate the applicability of the proposed framework, the optimized ANN was deployed as a web-based environmental sound classification application called UrbanSoundAI. The deployment was implemented using the Flask web framework and maintained the same preprocessing procedure used while developing the model. Users upload WAV audio files through a graphical web interface. The backend extracts 40 MFCC coefficients using the Librosa library after validating the uploaded file, [4], [15], [16]. The mean and standard deviation of the extracted coefficients are computed to generate an 80-dimensional feature vector, which is then normalized using the previously fitted StandardScaler before being fed into the trained ANN model for environmental sound classification [5].

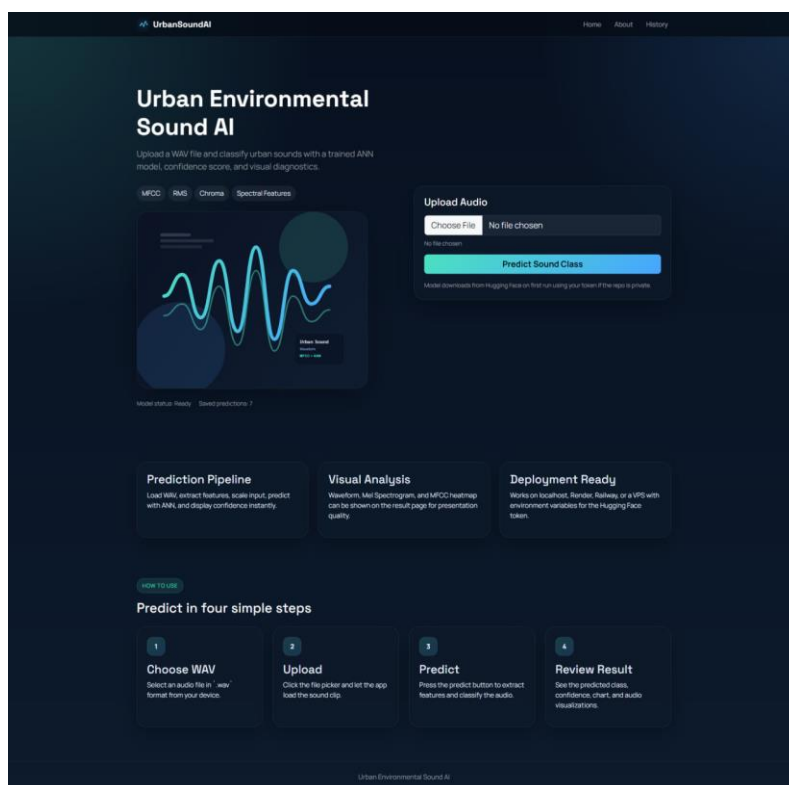


Figure 4. Home Interface of the UrbanSoundAI Web Application.

The application also shows the prediction confidence along with the predicted sound category. The prediction results are synchronized with the visualizations of the waveform, Mel spectrogram and MFCC heatmap, assist the user in interpreting the model output. Moreover, the system maintains a history of predictions which users can review at any time.

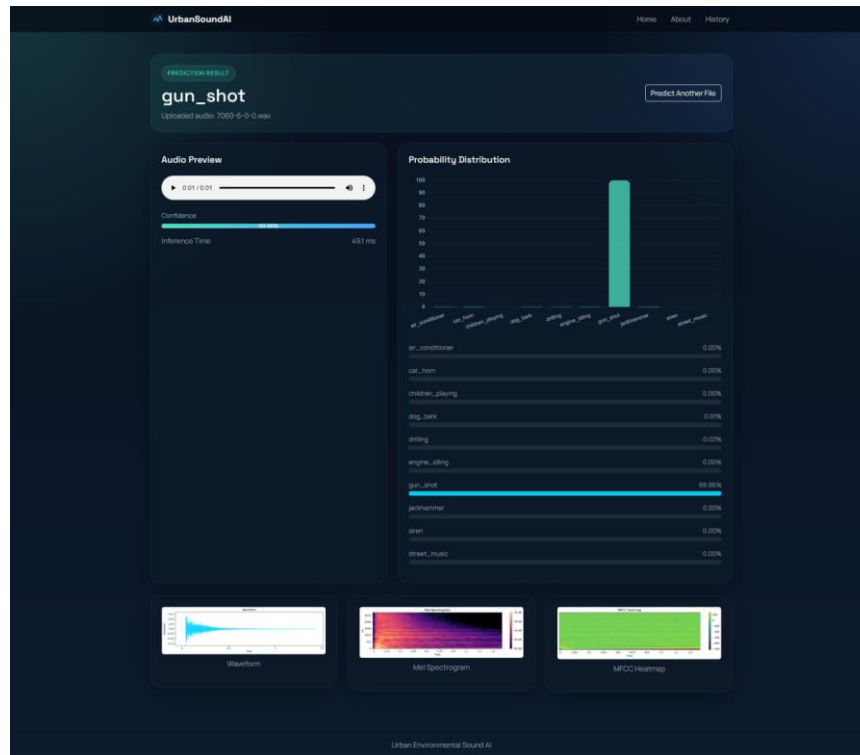


Figure 5. Prediction Result Generated by UrbanSoundAI

The deployment demonstrates that the proposed framework is suitable for both offline experimentation and practical real-time environmental sound recognition.

V.CONCLUSION

This paper presented a comparative evaluation of an optimized Artificial Neural Network(ANN) and a Support Vector Machine (SVM) for environmental sound classification using the UrbanSound8K benchmark dataset. An MFCC-based feature extraction framework was developed, in which the mean and standard deviation of 40 Mel-Frequency Cepstral Coefficients (MFCCs) were concatenated to construct an 80-dimensional acoustic feature vector. Prior to model training, feature normalization was performed using the StandardScaler method to ensure consistent feature scaling. This feature representation follows the typical pattern recognition approach of feature selection and standardization, which have been shown to be important for better classification accuracy [10]. Experimental results showed that the optimized network using the ANN gives the best classification accuracy of 94.81%, a slightly higher than the SVM's classification accuracy of 94.66%. Furthermore, the optimized ANN model was successfully deployed as the UrbanSoundAI web application, enabling near real-time environmental sound classification while providing waveform, Mel spectrogram, and MFCC heatmap visualizations for prediction interpretation. The deployment results demonstrate the practical feasibility of the proposed framework for real-world environmental sound recognition applications.

Overall, the suggested framework provides an accurate, reproducible and robust solution for environmental sound classification by integrating MFCC-based feature extraction with optimized machine learning and deep learning techniques. Future research may focus on evaluating the proposed framework using more large and diverse datasets, investigating advanced deep learning architectures, incorporating explainable artificial intelligence (XAI) methods to improve model interpretability, also optimizing the system for deployment on mobile and edge computing platforms.

REFERENCES

1. J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research, " in Proc. 22nd ACM Int. Conf. Multimedia (MM), Orlando, FL, USA, Nov. 2014, pp. 1041–1044.
2. K. J. Piczak, "Environmental sound classification with convolutional neural networks, " in Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP), Boston, MA, USA, Sept. 2015, pp. 1–6.
3. Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network, " in Proc. IEEE Int. Conf. Acoust. , Speech Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 2721–2725.
4. B. McFee et al. , "librosa: Audio and music signal analysis in Python, " in Proc. 14th Python Sci. Conf. (SciPy), Austin, TX, USA, Jul. 2015, pp. 18–25.
5. F. Pedregosa et al. , "Scikit-learn: Machine learning in Python, " J. Mach. Learn. Res. , vol. 12, pp. 2825–2830, 2011.

6. M. Abadi et al. , "TensorFlow: A System for Large-Scale Machine Learning, " in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI), Savannah, GA, USA, Nov. 2016, pp. 265–283.
7. F. Chollet et al. , Keras, 2015. [Online]. Available: <https://github.com/keras-team/keras>
8. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
9. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
10. C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
11. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.
12. C. Cortes and V. Vapnik, "Support-vector networks, " Mach. Learn. , vol. 20, no. 3, pp. 273–297, 1995.
13. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization, " in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015.
14. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization, " J. Mach. Learn. Res. , vol. 13, pp. 281–305, 2012.
15. D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab, " Columbia University, New York, NY, USA, Tech. Rep. , 2005.
16. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, " IEEE Trans. Acoust. , Speech, Signal Process. , vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
17. M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition, " Pattern Recognit. Lett. , vol. 24, no. 15, pp. 2895–2907, Nov. 2003.