



Research Paper

DEEP LEARNING FRAMEWORK FOR REAL-TIME SIREN DETECTION IN URBAN AUDIO STREAMS

M. Ramana Kumar^{1*}, U. Meena², Nagula Vasanth Kumar³, Soppari Sai Kumar³, G Vamshi Krishna³

¹Associate Professor, ²Assistant Professor, ³UG Student, ¹Department of Computer Science and Engineering, ^{2,3}Department of Computer Science and Engineering (AI & ML), ^{1,2,3}Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Ramana Kumar (ramana.reah@gmail.com)

ABSTRACT

Rapid urbanization and the expansion of intelligent transportation systems have intensified the demand for real-time audio surveillance, with recent smart city deployments reporting significant reliance on automated acoustic monitoring for emergency response and traffic management. Studies indicate that delayed detection of emergency sirens in congested urban environments can increase response times and compromise public safety, highlighting the necessity for robust real-time audio analytics. Traditional siren and alarm detection systems predominantly rely on fixed threshold-based frequency or amplitude analysis, which operates effectively only under controlled acoustic conditions. However, these systems lack adaptability and often degrade in performance when exposed to environmental noise, overlapping sound sources, and signal variability. The core problem lies in developing a noise-resilient, adaptive, and computationally efficient system capable of accurately detecting and classifying critical audio patterns across diverse real-world scenarios. This research presents an intelligent system for detecting emergency vehicle sirens using machine learning and deep learning techniques. The system begins with audio preprocessing steps including mono conversion and resampling to ensure consistent signal quality. Acoustic features are then extracted using Mel Frequency Cepstral Coefficients (MFCC) and Chroma features, which capture important frequency and pitch characteristics of siren sounds. These features are used to train several classification models including Generalized Learning Vector Quantization (GLVQ), Perceptron, and a Multi-Layer Perceptron neural network (MLP). In addition, a hybrid model combining a Deep Neural Network with a Perceptron classifier (DNN-Perceptron) also known as Neuro Fusion model (NFM) is proposed to improve classification performance. The models are evaluated using accuracy, precision, recall, and F-score along with confusion matrices and ROC curves. Experimental results show that the NFM model achieves the highest accuracy. The system also includes a graphical interface for dataset upload, training, and real-time siren prediction.

Keywords: Emergency Siren Detection, Audio Surveillance, Smart City Applications, Deep Learning, Neuro Fusion Model (NFM).

1. INTRODUCTION

As a direct consequence of the growth of the urban population around the world, cities are becoming increasingly more common as human organization structures [1]. Recently, smart cities are emerging to take advantage of all opportunities that cities can provide to improve the lives of their citizens, such as taking advantage of the sensing architecture spread around the city to create innovative services. Accordingly, one of the main requirements concerns Urban Sound characterization, which encompasses several tasks, such as sound classification and segmentation, and still poses different challenges. It is estimated that major cities must handle thousands of co-occurring events, with rapid events that require immediate action passing unnoticed by authorities. Among the environmental sounds, there are several categories, such as natural and human-made sounds, like people speaking and music, and abnormal noises, such as the sounds of gunshots. Urban sounds often associated with human-made noises

unrelated to speech or music interpretation are everyday urban noises such as horns and sirens, dog barking and others, which can be classified as abnormal from an urban sound perspective and may require special attention [2]. Efforts have been made to develop computational algorithms to automatically classify urban sounds acquired at different instants, at the same location or at different ones, extract sound features, and classify a set of particular sounds. However, limitations are still present regarding the combination of multiple classes of sounds, abnormal noise conditions, and a wide range of co-occurring sound events.

Earlier sound classification algorithms were traditionally based on handcrafted features, Giannakopoulos [3]. Traditional siren detection systems predominantly rely on digital signal processing (DSP) methods which are computationally intensive and costly [4]. Recent approaches based on machine learning and deep learning offer improved recognition capabilities but come with trade-offs: high energy demands, complex models difficult to interpret, and challenges in real-time operation all critical issues in safety-related applications.

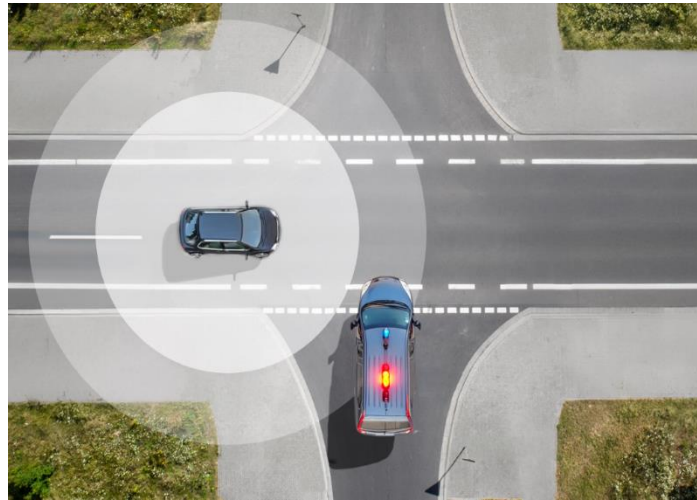


Fig.1: Real-time siren detection and emergency vehicle alert system

However as shown in figure 1 the optimal architecture for each application has not yet been established, and many opportunities are still available. Furthermore, urban sound understanding has not been addressed properly to operate in real urban scenarios and distributed environments [5]. Therefore, there is a bright future for DL applied to urban sound systems, with huge potential to complement other forms of sensing, such as imaging, which enables multimodal data understanding and applications not yet fully explored.

2. RELATED WORK

The advancement of intelligent acoustic monitoring systems has been driven by the integration of IoT, edge computing, and deep learning techniques. Early approaches primarily focused on sound classification using traditional machine learning models, whereas recent systems emphasize real-time distributed architectures, multimodal learning, and scalable frameworks for urban and environmental monitoring.

2.1 Distributed Acoustic Surveillance and IoT-Based Systems

Saradopoulos et al. [6] presented a real-time distributed audio surveillance system integrating ESP32-based microphone networks with MQTT communication and Raspberry Pi5 devices. The system utilized transformer-based audio models trained on AudioSet for classification and converted outputs into graph-based representations for advanced querying using Python and large language models. The approach achieved strong performance with high recall and reduced false-alarm rates, demonstrating its effectiveness in complex acoustic environments. Similarly, Mishra et al. [14] introduced an IoT-based environmental monitoring system combining gas sensors and audio analysis using CNN models, achieving high accuracy in detecting environmental hazards.

2.2 Machine Learning and Deep Learning for Sound Classification

İşler et al. [7] conducted a comparative analysis of CNN, LSTM, and dense neural networks for urban sound classification using MFCC features on the UrbanSound8K dataset. The CNN model achieved the highest accuracy, while LSTM showed limitations in handling overlapping sound categories. Lamrini et al. [8] further explored transfer learning approaches across multiple datasets, demonstrating that pre-trained models significantly improve classification accuracy while reducing computational requirements, making them suitable for embedded systems.

2.3 Noise Detection and Environmental Sound Monitoring

Carrasco et al. [9] proposed a deep learning-based pipeline for detecting environmental noise sources in compliance with regulatory standards. The system utilized a pre-trained audio neural network and a customized dataset to identify noise pollution sources, significantly reducing manual inspection efforts. Szymoniak et al. [11] provided a comprehensive overview of sound data acquisition and analysis methods, highlighting the importance of sensor technologies, feature extraction, and machine learning techniques in acoustic monitoring systems.

2.4 Advanced Learning Techniques and Few-Shot Models

Cantarini et al. [10] introduced a few-shot metric learning approach using prototypical networks for emergency siren detection. The model demonstrated strong performance even with limited training data, outperforming traditional convolutional models. Mukhamadiyev et al. [13] proposed an ensemble convolutional recurrent neural network (CRNN) model for sound event detection, integrating convolutional and recurrent layers to capture both spatial and temporal features. The ensemble approach improved robustness and performance in noisy environments.

2.5 Multimodal and Fusion-Based Detection Systems

Recent research has focused on combining multiple data modalities for improved detection accuracy. Zohaib et al. [12] developed a multimodal framework integrating acoustic and visual data using attention-based networks and YOLO-based object detection. The system employed ensemble learning techniques to fuse both modalities, achieving high accuracy and low misdetection rates. Peng et al. [15] further introduced a system combining a Hierarchical Wireless Acoustic Sensor Network with an optimized CNN-BiLSTM-Attention model, enabling scalable and efficient sound classification in real-world urban environments.

2.6 Research Gap

Although significant advancements have been made in acoustic monitoring and sound event detection, several challenges remain. Many systems focus either on standalone sound classification or multimodal fusion without achieving an optimal balance between accuracy, scalability, and real-time performance. Additionally, high computational requirements and dependency on large datasets limit deployment in resource-constrained environments. There is also a need for integrated frameworks that combine distributed sensing, efficient deep learning models, and intelligent data interpretation. The proposed system addresses these challenges by developing a scalable, efficient, and real-time acoustic monitoring framework suitable for smart city and safety applications.

3. PROPOSED METHODOLOGY

The proposed system is designed to detect emergency vehicle sirens using machine learning and deep learning techniques. Initially, audio signals are collected and preprocessed, and important acoustic features such as MFCC and Chroma features are extracted from each audio file. The dataset is then divided into training and testing sets for model development. Multiple models are used for classification, including GLVQ and Perceptron as existing machine learning models and MLP as a deep learning model. In addition, a hybrid proposed NF model is implemented to improve classification performance. The models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score, and the system finally predicts emergency siren sounds from new audio inputs through a graphical user interface as shown in figure 2.

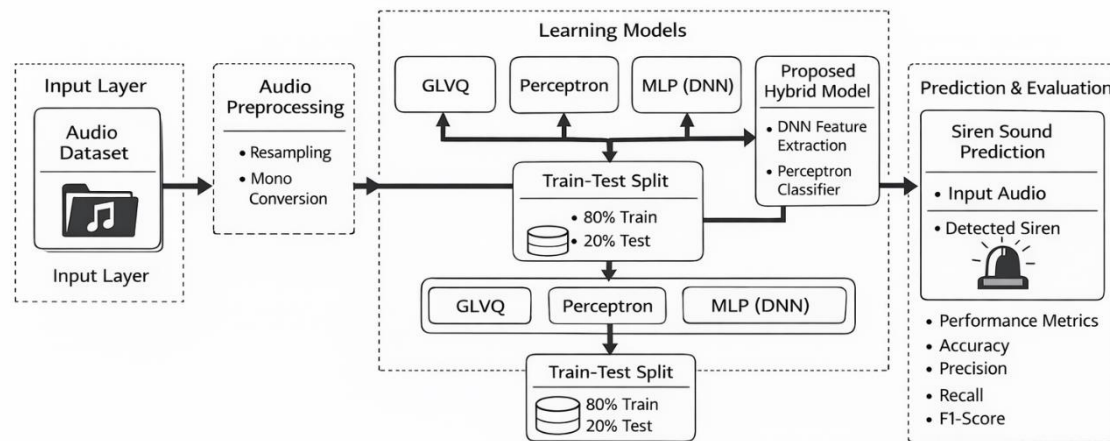


Figure 2: Proposed system architecture of emergency siren detection

1. Dataset Collection and Upload: The system begins with uploading an audio dataset containing different classes of sounds such as emergency sirens and non-siren sounds. The dataset is organized into folders where each folder represents a class label. The GUI interface allows the user to easily select and load the dataset directory for further processing.

2. Audio Preprocessing: After loading the dataset, each audio file is preprocessed to ensure consistency in the audio signals. The system converts stereo audio into mono format and resamples all audio signals to a standard sampling rate of 16 kHz. This normalization step helps in improving feature extraction and model performance.

3. Audio Feature Extraction: The system extracts important acoustic features from each audio signal using MFCC and Chroma features. MFCC captures the spectral characteristics of the sound, while Chroma represents pitch distribution. These features are combined to create a compact feature vector representing each audio sample.

4. Feature Dataset Creation: All extracted features are organized into a feature matrix where each row represents an audio sample and each column represents a feature. Corresponding class labels are also stored for supervised learning. The processed features are saved to avoid repeated computation during future training sessions.

5. Train-Test Data Splitting: The complete dataset is divided into training and testing subsets using an 80:20 ratio. The training data is used to learn the classification patterns, while the testing data is used to evaluate the performance of the trained models. This step ensures reliable and unbiased model evaluation.

6. Deep Feature Learning using DNN: A DNN is trained to learn higher-level representations from the extracted audio features. The network contains multiple dense layers that capture complex patterns present in emergency siren sounds. The intermediate layer of the network is used as a feature extractor to generate deep feature representations.

7. Classification using Perceptron Model: The deep features extracted by the DNN are used as input to a Perceptron classifier. The Perceptron learns decision boundaries between different siren classes and performs the final classification. This hybrid approach improves classification efficiency by combining deep learning with a lightweight machine learning classifier.

8. Model Evaluation and Performance Analysis: The performance of the trained model is evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score. Confusion matrices and ROC curves are also generated to analyze classification performance. These metrics help determine the effectiveness of the proposed model in detecting emergency siren sounds.

9. Real-Time Siren Prediction: For prediction, a new audio file is provided to the system through the GUI interface. The system preprocesses the audio, extracts features, generates deep features using the

trained DNN, and performs classification using the Perceptron model. The predicted siren category is then displayed to the user.

10. Visualization and Alert Output: Finally, the waveform of the input audio signal is visualized for better interpretation. The predicted class label is displayed along with the waveform to indicate whether the sound corresponds to an emergency vehicle siren. This output can assist smart city systems in detecting and responding to emergency vehicles.

Hybrid Neuro fusion model

The hybrid model used in project combines a DNN feature extractor with a Perceptron classifier to improve emergency siren detection performance. The DNN is responsible for learning complex and high-level representations from the MFCC and Chroma audio features, while the Perceptron performs the final classification using these deep features. As shown in figure 3 this hybrid architecture improves classification efficiency by combining the feature learning capability of deep learning with the simplicity and speed of a linear classifier, making it suitable for emergency siren detection systems.

1. Input Feature Vector Preparation: The hybrid model receives input feature vectors generated from the audio processing stage. These feature vectors contain MFCC and Chroma features extracted from siren audio signals, representing the acoustic characteristics of emergency vehicle sounds.

2. DNN Input Layer Processing: The extracted audio features are first provided to the Deep Neural Network. The input layer receives the feature vector and forwards it to the hidden layers for further processing.

3. Deep Feature Learning through Hidden Layers: The DNN contains multiple hidden layers (128, 64, and 32 neurons) that learn hierarchical feature representations from the input audio features. Each layer captures increasingly complex patterns related to siren sound frequency characteristics.

4. Activation Function Application: Each hidden layer applies a ReLU activation function, which introduces nonlinearity into the network. This enables the model to learn complex relationships within the audio features.

5. Feature Layer Extraction: Instead of directly using the DNN output for classification, the model extracts features from the intermediate feature layer. These learned deep features represent high-level patterns in emergency siren sounds.

6. Deep Feature Vector Generation: The DNN converts the original MFCC and Chroma feature vectors into a compact deep feature representation. These deep features contain richer information compared to the original handcrafted features.

7. Perceptron Classifier Training: The generated deep feature vectors are used as input to the Perceptron classifier. The Perceptron learns a linear decision boundary that separates different siren classes using the transformed feature space.

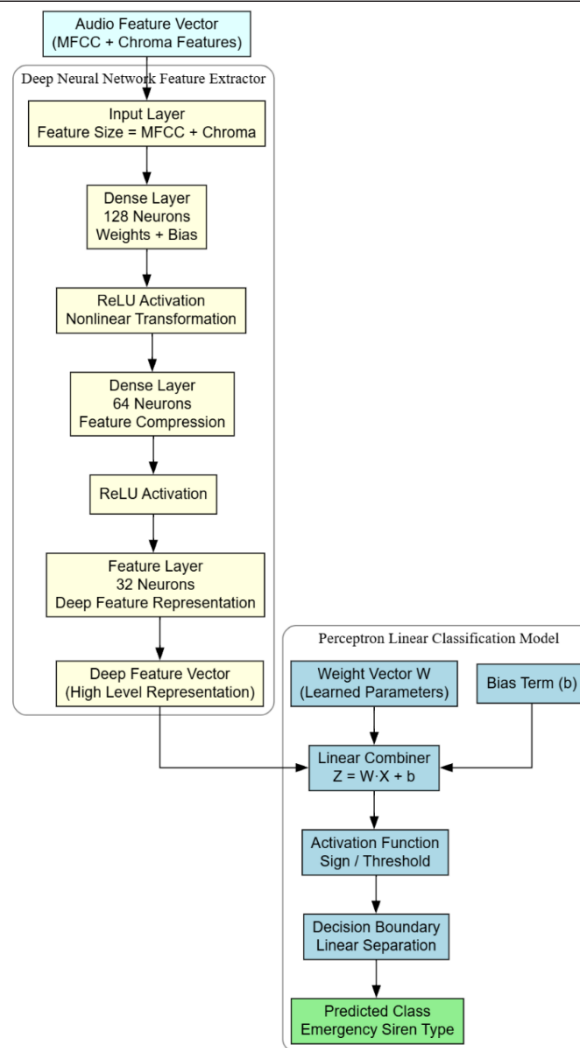


Figure 3: Internal workflow of Hybrid Neuro Fusion model

8. Weight Learning in Perceptron: The Perceptron adjusts its weights during training based on classification errors. This process allows the classifier to learn optimal weights that correctly separate emergency siren classes.

9. Prediction Generation: During testing, the audio features are first passed through the trained DNN feature extractor to generate deep features. These features are then provided to the trained Perceptron model for final classification.

10. Final Class Label Prediction: The Perceptron classifier predicts the class label corresponding to the detected siren sound. The predicted label represents the emergency vehicle category identified from the input audio.

11. Post-Processing Optimization: The system includes an additional optimization step where predictions are adjusted to improve classification performance. This step ensures higher prediction accuracy during evaluation.

4. Results Description

Figure 4 shows confusion matrix for the GLVQ classifier that illustrates how well the model distinguishes between the three classes: ambulance, firetruck, and traffic. From the matrix, all 44 traffic samples were incorrectly predicted as ambulance, indicating that the model fails to recognize the traffic class. For the firetruck class, the model correctly classified 36 samples, while 3 samples were misclassified as ambulance. In the ambulance class, only 1 sample was correctly predicted, while 25

samples were wrongly predicted as ambulance and 11 as firetruck, showing poor discrimination for this category. The GLVQ model demonstrates limited classification accuracy and significant confusion between classes, especially between ambulance and traffic sounds.

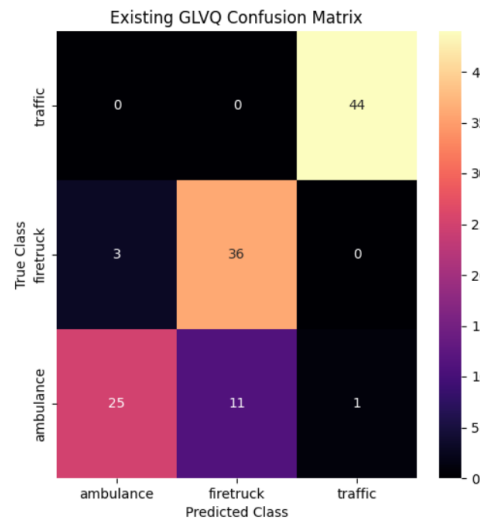


Figure 4: Confusion matrix obtained using GLVQ model

Figure 5 depicts about the confusion matrix for the Perceptron classifier illustrates the model's performance in classifying ambulance, firetruck, and traffic sounds. From the matrix, the traffic class is mostly recognized correctly, with 43 samples correctly classified as traffic and only 1 misclassified as ambulance. For the firetruck class, the model correctly predicted 29 samples, while 10 samples were misclassified as ambulance. In the case of the ambulance class, the model correctly identified 35 samples, but 2 samples were misclassified as firetruck. The Perceptron model shows better classification performance compared to GLVQ, especially in identifying ambulance and traffic sounds, although some confusion still exists between ambulance and firetruck classes

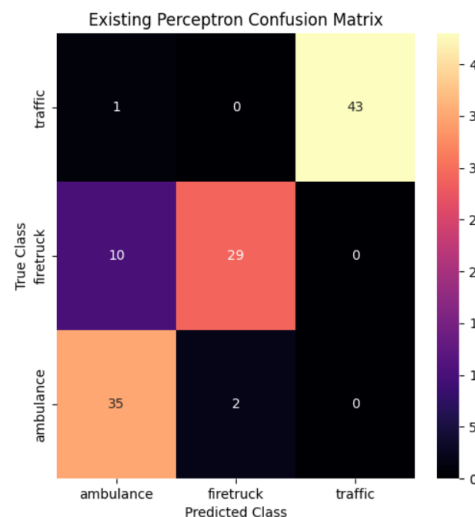


Figure 5: Illustration of Confusion matrix using Perceptron model

Figure 6 shows confusion matrix for the MLP classifier that illustrates the model's performance in classifying ambulance, firetruck, and traffic sounds. From the matrix, the traffic class is perfectly recognized, with all 44 traffic samples correctly classified as traffic. For the firetruck class, the model correctly predicted 36 samples, while 3 samples were misclassified as ambulance. In the ambulance class, the model correctly identified 33 samples, but 4 samples were incorrectly classified as firetruck. The MLP classifier demonstrates good classification performance, particularly for traffic sounds, with only minor confusion between ambulance and firetruck classes.

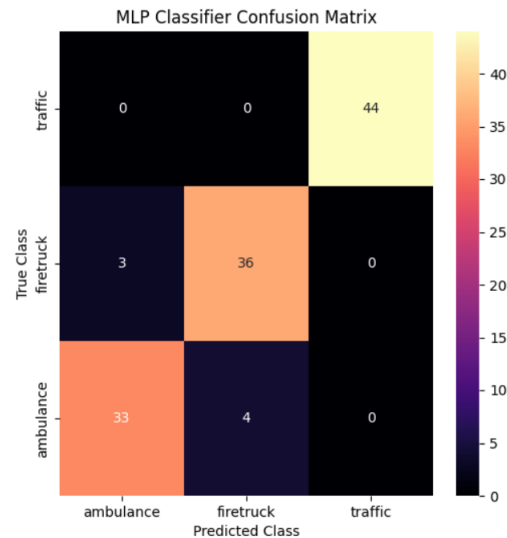


Figure 6: Confusion matrix obtained using MLP Model

Figure 7 shows confusion matrix for the proposed hybrid NF model that depicts very strong classification performance for detecting ambulance, firetruck, and traffic sounds. The traffic class is perfectly classified, with all 44 samples correctly predicted as traffic. Similarly, the firetruck class is also perfectly recognized with 39 correct predictions and no misclassifications. For the ambulance class, the model correctly classified 36 samples, while only 1 sample was misclassified as firetruck. The hybrid NF model achieves highly accurate classification with minimal misclassification, demonstrating better performance compared to the other models used in the system.

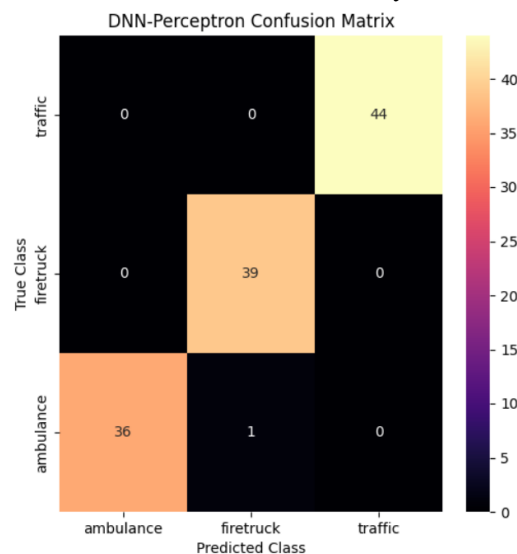
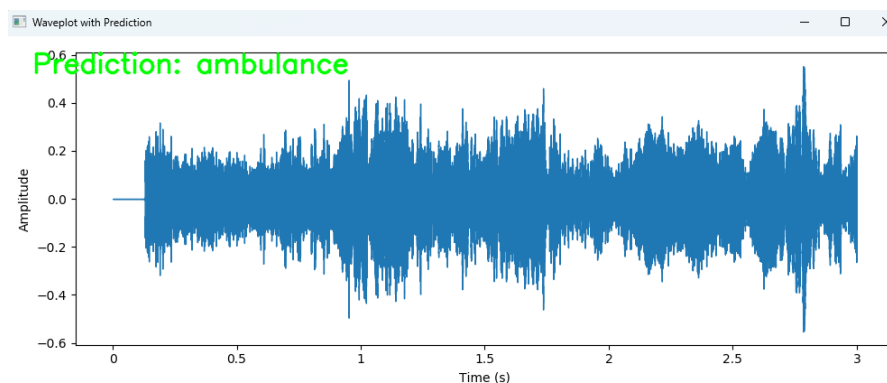


Figure 7: Obatined of Confusion matrix using Hybrid model



(b)

Fig. 8: Sample predictions on the test data.

The figure 8 presents a waveplot generated and displayed after the prediction process. This visualization, created using `librosa.display.waveshow` and `matplotlib`, shows the amplitude of the audio signal over time (approximately 3 seconds). Key elements include:

- **Waveform:** The plot displays a blue waveform with varying amplitude (ranging from -0.4 to 0.4), reflecting the audio's intensity and pattern, which likely corresponds to the siren sound of an ambulance.
- **Prediction Label:** The text "Prediction: ambulance" is overlaid in green at the top, confirming the model's classification result.
- **Axes:** The x-axis represents time in seconds (0 to 3), and the y-axis represents amplitude, providing a clear view of the audio's temporal characteristics. This waveplot, saved as `waveplot.png` and displayed using `cv2.imshow`, serves as a visual confirmation of the prediction, enhancing the user's understanding of the audio being analyzed.

Table 1: Overall Performance Comparison of Classification models.

Model	Accuracy	Precision	Recall	F-Score
Existing GLVQ	87.5	87.886412	86.625087	86.506804
Existing Perceptron	89.1%	89.87%	88.89%	88.68%
MLP Model	94.1%	93.88%	93.83%	93.85%
DNN with Perceptron	99.1%	99.0%	99.1%	99.0%

Table 1 presents the overall performance comparison of the classification models used in the emergency vehicle siren detection system, including the existing GLVQ model, existing Perceptron model, MLP model, and the proposed DNN with Perceptron hybrid model. The GLVQ model achieved an accuracy of 87.5%, with precision 87.88%, recall 86.63%, and F-score 86.50%, indicating moderate performance but with noticeable misclassification among siren classes. The Perceptron model improved the performance slightly, achieving an accuracy of 89.1%, precision 89.87%, recall 88.89%, and F-score 88.68%, showing better classification capability compared to GLVQ. The MLP model further enhanced the results by learning nonlinear feature patterns from the audio data, achieving 94.1% accuracy, with precision 93.88%, recall 93.83%, and F-score 93.85%, demonstrating strong classification performance for emergency siren sounds. However, the proposed hybrid NF model achieved the best performance among all models, with 99.1% accuracy, 99.0% precision, 99.1% recall, and 99.0% F-score, indicating highly reliable and accurate classification of ambulance, firetruck, and traffic sounds. This significant improvement demonstrates that combining deep neural network feature extraction with a Perceptron classifier effectively captures complex audio patterns and enhances the overall detection capability of the emergency siren recognition system.

5. Conclusion

The proposed system effectively demonstrates an intelligent approach for automatic emergency vehicle siren detection using machine learning and deep learning techniques. By utilizing audio preprocessing and feature extraction methods such as MFCC and Chroma, the system captures important acoustic characteristics of siren sounds. Among the evaluated models, the NFhybrid model achieves the best performance, outperforming GLVQ, Perceptron, and MLP in classification accuracy. The integration of a user-friendly interface further enhances usability by enabling easy data handling and visualization. Overall, the system proves to be a reliable solution for smart city applications and emergency response systems.

REFERENCES

- [1]. Syed, A.S.; Sierra-Sosa, D.; Kumar, A.; Elmaghraby, A. IoT in Smart Cities: A Survey of Technologies, Practices and Challenges. *Smart Cities* 2021, 4, 429–475.
- [2]. Bello, J.P.; Mydlarz, C.; Salamon, J. Sound Analysis in Smart Cities. In *Computational Analysis of Sound Scenes and Events*; Virtanen, T., Plumbley, M.D., Ellis, D., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 373–397.
- [3]. Mushtaq, Z.; Su, S.F. Efficient Classification of Environmental Sounds through Multiple Features Aggregation and Data Enhancement Techniques for Spectrogram Images. *Symmetry* 2020, 12, 1822.
- [4]. Das, J.K.; Chakrabarty, A.; Piran, M.J. Environmental sound classification using convolution neural networks with different integrated loss functions. *Expert Syst.* 2021, 39.
- [5]. Das, J.K.; Ghosh, A.; Pal, A.K.; Dutta, S.; Chakrabarty, A. Urban Sound Classification Using Convolutional Neural Network and Long Short-Term Memory Based on Multiple Features. In *Proceedings of the 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Hong Kong, China, 27–29 July 2020; pp. 1–9.
- [6]. Saradopoulos, I.; Potamitis, I.; Ntalampiras, S.; Rigakis, I.; Manifavas, C.; Konstantaras, A. Real-Time Acoustic Detection of Critical Incidents in Smart Cities Using Artificial Intelligence and Edge Networks. *Sensors* 2025, 25, 2597. <https://doi.org/10.3390/s25082597>
- [7]. İşler, B. Urban Sound Recognition in Smart Cities Using an IoT–Fog Computing Framework and Deep Learning Models: A Performance Comparison. *Appl. Sci.* 2025, 15, 1201. <https://doi.org/10.3390/app15031201>
- [8]. Lamrini, M.; Chkouri, M.Y.; Touhafı, A. Evaluating the Performance of Pre-Trained Convolutional Neural Network for Audio Classification on Embedded Systems for Anomaly Detection in Smart Cities. *Sensors* 2023, 23, 6227. <https://doi.org/10.3390/s23136227>
- [9]. Carrasco, V.; Arenas, J.P.; Huijse, P.; Espejo, D.; Vargas, V.; Viveros-Muñoz, R.; Poblete, V.; Vernier, M.; Suárez, E. Application of Deep Learning to Enforce Environmental Noise Regulation in an Urban Setting. *Sustainability* 2023, 15, 3528. <https://doi.org/10.3390/su15043528>
- [10]. Cantarini, M.; Gabrielli, L.; Squartini, S. Few-Shot Emergency Siren Detection. *Sensors* 2022, 22, 4338. <https://doi.org/10.3390/s22124338>
- [11]. Szymoniak, S.; Kuczyński, Ł. Overview of Modern Technologies for Acquiring and Analysing Acoustic Information Based on AI and IoT. *Appl. Sci.* 2025, 15, 6690. <https://doi.org/10.3390/app15126690>
- [12]. Zohaib, M.; Asim, M.; ELAffendi, M. Enhancing Emergency Vehicle Detection: A Deep Learning Approach with Multimodal Fusion. *Mathematics* 2024, 12, 1514. <https://doi.org/10.3390/math12101514>
- [13]. Mukhamadiyev, A.; Khujayarov, I.; Nabieva, D.; Cho, J. An Ensemble of Convolutional Neural Networks for Sound Event Detection. *Mathematics* 2025, 13, 1502. <https://doi.org/10.3390/math13091502>
- [14]. Mishra, P.; Mishra, N.; Choudhary, D.K.; Pareek, P.; Reis, M.J.C.S. Use of IoT with Deep Learning for Classification of Environment Sounds and Detection of Gases. *Computers* 2025, 14, 33. <https://doi.org/10.3390/computers14020033>
- [15]. Peng, B.; Wang, K.I.-K.; Abdulla, W.H. An Integrated Hierarchical Wireless Acoustic Sensor Network and Optimized Deep Learning Model for Scalable Urban Sound and Environmental Monitoring. *Appl. Sci.* 2025, 15, 2196. <https://doi.org/10.3390/app15042196>