International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

End-to-End CNN-RNN Model for Robust Speech Command Recognition

P Mahipal Reddy[1*], Kotha Harshitha[2], Palithepu Rajkumar[2], Puli Purnathejitha[2], Bitla Saketh[2]

[1]Associate Professor, [1,2]Department of Computer Science and Engineering (Data science)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

*Correspondence: P Ma-hipal Reddy (mahipalreddy.pulyala@vecw.edu.in)

ABSTRACT

Environmental audio signals contain rich spectral and temporal characteristics that enable automatic classification of diverse sound events. In real-world scenarios, accurate identification of sounds such as human activities, falls, abnormal environmental noises, machine faults, and acoustic anomalies is essential for applications in safety monitoring, healthcare, surveillance, and intelligent systems. Traditional machine learning methods, including K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Adaptive Boosting Classifier (AdaBoost), and Linear Discriminant Analysis (LDA), rely on handcrafted features and shallow architectures. While effective for simple datasets, they struggle with complex, noisy audio and fail to capture high-level temporal dependencies. Their limited generalization across varying acoustic environments highlights the need for more advanced approaches. To address these challenges, this work proposes a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM). The system extracts comprehensive features using Librosa, including MFCC, Mel Spectrogram, Chroma, Zero Crossing Rate (ZCR), Root Mean Square (RMS), Spectral Contrast, Bandwidth, Centroid, and Tonnetz. These features represent both frequency and temporal characteristics of audio signals. The CNN learns hierarchical spectral patterns, while the LSTM captures sequential and long-term dependencies. This combination enables effective understanding of both sound content and temporal evolution. Additionally, a Flask-based API supports real-time classification by allowing external systems to send audio inputs and receive predictions instantly. Experimental results show that the proposed model significantly improves accuracy, robustness, and generalization compared to traditional methods.

Keywords: Environmental Audio Classification, Deep Learning, Convolutional Neural Network, MFCC, Long Short-Term Memory.

1. INTRODUCTION

With the rapid advancement of deep learning technologies, speech recognition has become a widely used and essential component in modern computing systems. The increasing popularity of smart devices has significantly enhanced the demand for voice-based interactions, allowing users to communicate with machines in a more natural and convenient manner. As a fundamental form of human–computer interaction, Automatic Speech Recognition (ASR) enables systems to interpret and process spoken language efficiently as shown in Fig. 1. Since its emergence in the 1970s, ASR has remained a key research area within the field of machine learning, continuously evolving with improvements in computational power and advanced algorithms.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

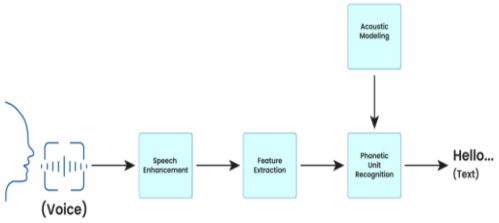


Fig 1: Speech command classification process.

Automatic Speech Recognition (ASR) refers to the process of converting spoken language into text by identifying and interpreting speech signals. An ASR system estimates the most probable sequence of words from a given audio input. Over the past few decades, ASR has become a significant area of research, as it provides an efficient medium for both human-to-human and human-to-machine communication. With continuous advancements, the technology has matured and is now seamlessly integrated into modern smart devices, increasing its adoption across various applications.

Applications such as Google Assistant, Amazon Alexa, and Apple Siri have transformed the way users interact with devices through voice commands. This growing trend is driven by several factors, including improved computational power, availability of large-scale datasets through big data technologies, and the rapid increase in smart device usage such as smartphones, wearables, smart home systems, and in-vehicle infotainment systems. To enhance user experience, it is essential to develop robust and efficient speech-based interfaces that enable natural and intuitive interaction, regardless of users' background or technical knowledge.

2. Related Work

The field of environmental audio and speech recognition has evolved significantly with the transition from traditional machine learning methods to deep learning-based architectures. Early approaches relied on handcrafted features and shallow classifiers, which showed limitations in handling noisy and complex real-world audio data. Recent advancements focus on hybrid deep learning models, multimodal learning, and noise-robust techniques to improve performance and generalization.

2.1 Feature Engineering and Representation Learning

Effective feature extraction plays a crucial role in audio classification systems. Begazo et al. [1] explored the combination of spectral features and spectrogram images, demonstrating that feature fusion significantly improves classification accuracy. Similarly, Kubanek et al. [3] proposed a novel representation method using multiple convolution layers across time and frequency domains, treating audio signals as RGB-like images to capture richer feature details.

Tamazin et al. [14] introduced an enhanced PNCC-based feature extraction technique using gammatone filtering, which significantly improves performance under low signal-to-noise ratio conditions. These studies highlight that advanced feature representations are essential for capturing both spectral and temporal characteristics of audio signals.

2.2 Deep Learning Architectures for Audio Classification

The adoption of deep learning models has greatly improved audio classification performance. Vujičić et al. [5] utilized a CNN-based architecture with Mel spectrogram inputs, achieving high accuracy and strong generalization.

Ouali et al. [10] further advanced this approach by combining CNN with Bidirectional LSTM, achieving superior accuracy through hybrid modeling. Similarly, Lin et al. [7] proposed a CNN model with phonetic posteriorgram features, outperforming traditional MFCC-based systems while reducing model complexity.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

2.3 Noise Robustness and Speech Enhancement

Handling noise remains a critical challenge in real-world audio applications. Ullah et al. [2] proposed a Convolutional Recurrent Network (CRN) for speech enhancement, improving both intelligibility and quality while maintaining computational efficiency.

Pervaiz et al. [4] addressed noise robustness through data augmentation techniques, demonstrating improved performance on noisy datasets. Additionally, Hussein et al. [8] incorporated noise removal and feature optimization in deep neural networks, achieving high accuracy with fast response time.

2.4 Multimodal and Advanced Recognition Systems

Recent research explores multimodal approaches to enhance recognition performance. Torrie et al. [6] introduced the MultiAVSR framework, combining audio and visual inputs with multi-task learning to improve robustness and reduce computational cost.

Jeon et al. [12] demonstrated that combining audio and visual features significantly boosts recognition accuracy compared to single-modality systems.

Yang et al. [13] further analyzed audiovisual models and highlighted challenges such as feature extraction and domain generalization, identifying advanced models like MoCo + Word2Vec as highly effective.

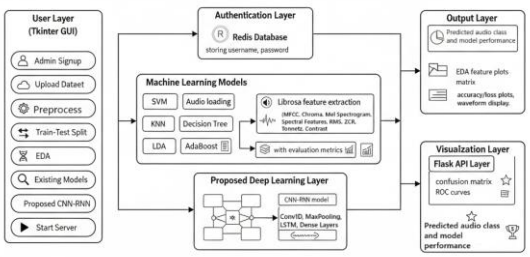
2.5 Applications and System-Level Implementations

Several studies have focused on real-world implementations of speech-based systems. Gupta et al. [11] developed a speech-controlled robotic system with low latency and high accuracy, demonstrating robustness in noisy environments.

Wang et al. [9] provided a comprehensive review of traditional and end-to-end speech recognition models, highlighting the shift toward deep learning-based approaches as the future of the field.

3. PROPOSED SYSTEM

The proposed system is designed to provide an accurate, automated, and intelligent solution for environmental sound classification using a hybrid deep learning architecture that integrates CNN and RNN. Unlike traditional approaches, where feature extraction and classification are separate processes, the proposed system follows an end-to-end learning approach capable of capturing complex patterns present in audio signals. This improves the system's ability to handle real-world challenges such as background noise, varying pitch levels, overlapping frequencies, and dynamic acoustic environments. The system starts with a comprehensive feature extraction process in which each audio signal is transformed into a rich set of descriptors using the Librosa library. These features include MFCC, Chroma, Mel Spectrogram, Spectral Contrast, Tonnetz, ZCR, RMS, and other spectral characteristics as shown in Fig. 2. Together, these features provide a detailed representation of both the frequency-domain and temporal aspects of the audio signal. By utilizing this diverse feature set, the system ensures meaningful and discriminative information is captured for accurate classification. The integration of CNN and RNN allows the model to learn both spatial and sequential patterns, resulting in improved accuracy, robustness, and generalization across different real-world audio scenarios.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig 2: Proposed system architecture

The system starts by allowing users to upload a structured dataset of audio recordings organized into class-specific folders, enabling efficient labeling and preprocessing. Each audio file is processed using Librosa to extract rich features such as MFCC, Chroma, Mel-Spectrogram, Spectral Contrast, Tonnetz, ZCR, RMS, Spectral Bandwidth, and Spectral Centroid, capturing spectral, tonal, and temporal characteristics. Exploratory Data Analysis (EDA) is then performed through visualizations like waveforms and feature maps to understand data patterns and detect anomalies. The data is standardized, shuffled, and split into training and testing sets to ensure balanced and unbiased learning. A hybrid CNN-RNN model is trained, where CNN layers learn spatial and frequency features while RNN layers capture temporal dependencies, enabling robust audio classification. The model is evaluated using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curves to assess performance. For real-time usage, users can input new audio files to receive instant predictions along with waveform visualizations. Finally, the system is deployed via a Flask-based API with Redis-backed authentication, and a user-friendly GUI allows seamless dataset management, model training, visualization, and prediction.

CNN–RNN Model

The CNN–RNN hybrid model internally processes audio features by learning spatial and temporal patterns simultaneously. The input audio feature sequence is first passed through convolutional layers, where multiple filters scan across the feature dimension to automatically learn local patterns such as frequency peaks, harmonic structures, and sudden spectral transitions. These convolutional responses are then down sampled using pooling layers to retain essential information while reducing dimensionality. The condensed feature maps are forwarded to a recurrent layer, such as LSTM, which models temporal dependencies and captures sequential changes across the audio signal. This helps the network understand how sound evolves over time. Finally, fully connected layers convert the learned representations into probability distributions over class labels using the soft max activation function. This combined architecture enables the model to extract robust hierarchical features ideal for sound classification as shown in Fig. 3.

Step 1: Input Feature Structuring and Normalization

The CNN–RNN pipeline begins by reshaping extracted audio features such as MFCCs, Mel spectrogram coefficients, chroma features, Tonnetz, and spectral descriptors into a consistent time–feature matrix. Each row corresponds to a specific time frame while each column represents a particular acoustic feature. Before feeding these into the network, normalization is applied to ensure that all feature values fall within a similar range. This avoids bias where large-valued features dominate the learning process. Proper structuring and normalization ensure stable gradient flow during training.

Step 2: Sequence Preparation and Dimensional Expansion

To make the input compatible with convolutional layers, the feature matrix is expanded with an additional channel dimension. This transformation converts the 2D array into a 3D tensor with dimensions reflecting time, feature count, and channel depth. The sequential arrangement preserves the chronological order of audio frames, enabling the network to model time dependencies effectively. This step ensures that the CNN filters can scan across meaningful acoustic regions. Maintaining sequence integrity is crucial for temporal modelling in later stages.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

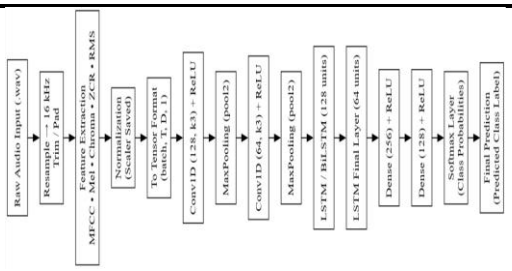


Fig 3: Internal workflow of Hybrid CNN-RNN model

Step 3: Convolutional Filtering (Conv1D Layer 1)

The first Conv1D layer applies multiple filters that slide along the time axis to extract local spectral patterns. Each filter is designed to detect specific patterns such as sudden energy bursts, harmonic transitions, frequency modulations, or characteristic sound edges. Convolution produces feature maps that highlight these local patterns across the audio timeline. This step allows the model to automatically learn low-level spectral features that manual feature engineering might miss. The output forms a rich representation of localized sound structures.

Step 4: Non-linear Activation (Re LU)

After convolution, a Rectified Linear Unit (Re LU) activation is applied to introduce non-linearity into the learning process. Re LU zeroes out all negative values, allowing only positive activations to pass through. This helps the network learn complex, non-linear relations between acoustic features and sound categories. It also reduces the vanishing gradient problem and accelerates convergence. Applying Re LU after convolution ensures that the model focuses on meaningful activations rather than noise.

Step 5: Convolutional Filtering (Conv1D Layer 2)

A second convolutional layer further processes the feature maps produced by the first layer. At this stage, filters learn more abstract and higher-level spectral representations by combining outputs from previous receptive fields. This allows the model to capture complex acoustic shapes such as multi-frequency interactions, resonance patterns, and composite sound signatures. The deeper convolution expands the network's capacity to learn intricate sound structures. These multi-level features provide a robust foundation for temporal analysis.

Step 6: Max Pooling for Dimensionality Reduction

Max Pooling is applied to reduce the temporal dimension of feature maps while preserving the most dominant activations. By selecting the maximum value within small windows, pooling compresses the data and filters out minor variations or noise. This step enhances computational efficiency and prevents overfitting by reducing feature redundancy. Pooling also introduces translation invariance, meaning small shifts in sound events do not affect recognition. The pooled maps contain the most informative spectral cues for subsequent analysis.

Step 7: Preparing Features for RNN/LSTM Layer

Before feeding pooled features into the LSTM, the feature maps are reshaped back into a sequential structure. Each time step is represented by a feature vector that summarizes information captured by CNN filters. This reshaping preserves the temporal ordering necessary for the RNN to identify patterns over time. It converts spatially-enhanced spectral features into a form suitable for time-series modelling. This bridging step integrates convolutional and recurrent learning into a unified pipeline.

Step 8: Temporal Modelling with LSTM

The LSTM layer processes the sequence of feature vectors frame by frame, learning how acoustic patterns evolve over time. LSTM cells maintain memory states that remember long-term dependencies, allowing the model to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

detect temporal dynamics such as gradual intensity changes, repetitive sound cycles, and characteristic rise-and-fall patterns. This temporal modelling is essential for distinguishing sounds that share spectral similarities but differ in timing. LSTM enhances the model's ability to classify complex, time-dependent audio events.

3. Result Description

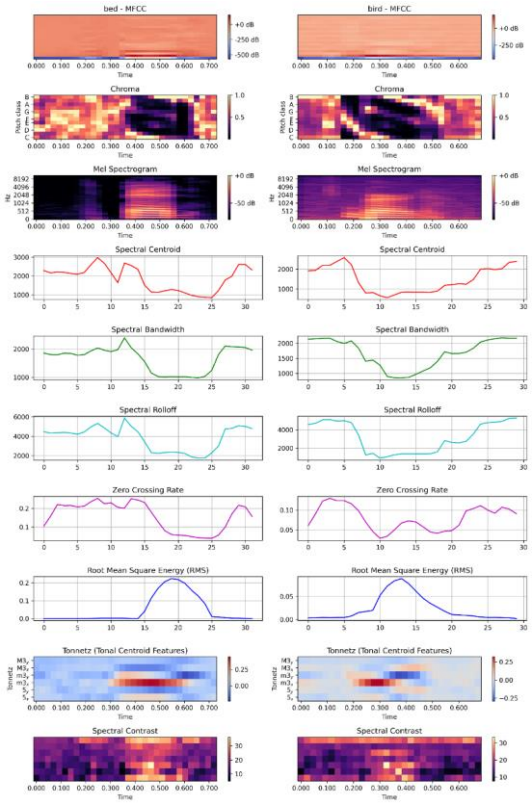


Fig. 4: Comparative EDA Feature Visualizations for Two Audio Classes

Fig. 4 presents the EDA results for two different audio classes bed and bird. For each class, multiple acoustic features are visualized to understand their frequency patterns, energy distribution, and temporal behavior. The plots include MFCC heatmaps, Chroma features, Mel-Spectrograms, Spectral Centroid curves, Spectral Bandwidth, Spectral Rolloff, ZCR, RMS Energy, Tonnetz (tonal centroid features), and Spectral Contrast. These visualizations help compare how different sound classes vary in terms of harmonic structure, amplitude changes, frequency emphasis, and spectral characteristics. Such insights support better feature engineering and contribute to improved performance of both machine learning and CNN–RNN models used in the system.

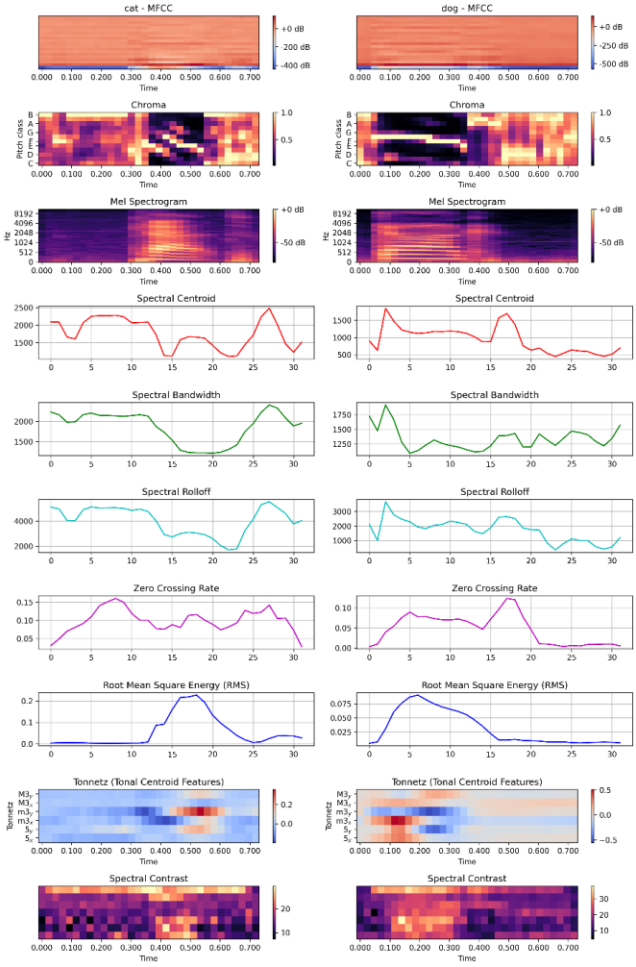


Fig. 5: EDA Feature Comparison Between "cat" and "dog" Audio Classes

Fig. 5 presents a comparative Exploratory Data Analysis (EDA) of the cat and dog audio classes. The visualizations include MFCC heatmaps, Chroma energy distributions, Mel-Spectrograms, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Zero-Crossing Rate, Root Mean Square (RMS) Energy, Tonnetz (tonal centroid features), and Spectral Contrast. These features illustrate how both animal sound categories differ in frequency content, harmonic patterns, amplitude variations, and temporal behaviour. The plots

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

reveal clear distinctions in spectral shape and energy spread, which help machine learning and CNN–RNN models learn class-specific acoustic patterns. This comparative analysis is essential for understanding feature relevance and improving model performance.

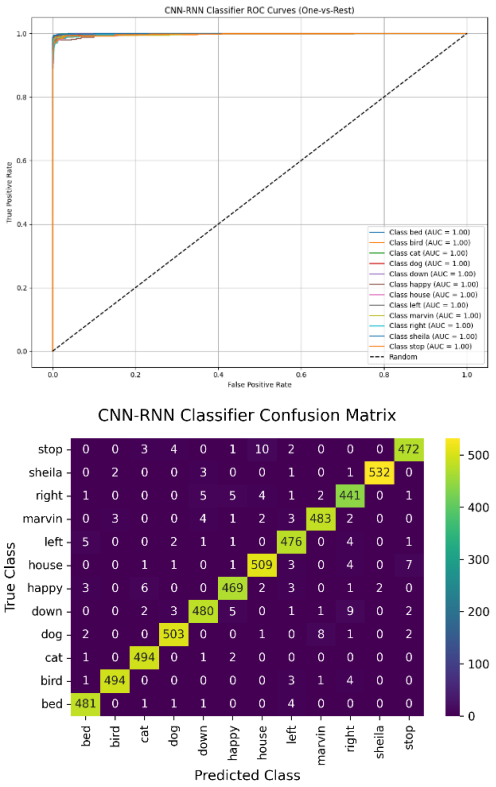


Fig. 6: Confusion matrix and ROC obtained of the Proposed CNN–RNN Classifier

Fig. 6 presents the confusion matrix and ROC of the proposed CNN–RNN hybrid deep learning model.The One-vs-Rest ROC Curves and the Confusion Matrix. The ROC Curve panel shows exceptionally strong classification performance across all 12 sound categories, with each class achieving an Area Under the Curve (AUC) of 1.00, indicating near-perfect separability and highly reliable prediction capability. The Confusion Matrix further validates this performance by showing an almost flawless diagonal structure, where the majority of samples are classified correctly with minimal misclassification. The CNN–RNN model captures both spectral and temporal dependencies in audio signals, enabling it to outperform traditional machine learning models significantly. Together, these results demonstrate the robustness, precision, and superior generalization ability of the proposed deep learning approach for sound signal classification.

Fig. 7: Flask Server Initialization for Real-Time Prediction

Fig. 7 shows the server initialization interface within the Sound Signal Classification System. When the “Start Server” button is activated, the system launches a Flask-based API server at the displayed address (http://0.0.0.0:5000). This server is specifically configured to run on Laptop-2, enabling real-time audio prediction requests from external devices or the Tkinter GUI. Once the server is active, users can send sound files to the API endpoint for processing, feature extraction, and classification using the trained CNN–RNN model. This functionality allows the system to operate as a distributed prediction service, supporting remote inference and multi-device integration.

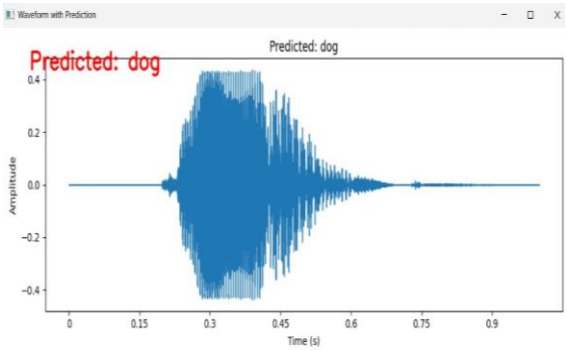


Fig. 8: Remote Prediction Output with Waveform Visualization

Fig. 8 shows the prediction interface of the Remote Prediction Client, which communicates with the Flask server running on Laptop-2. After selecting a test audio file, the system sends the file to the remote server, where the trained CNN–RNN model performs real-time feature extraction and classification. The

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

predicted class label is displayed both as text ("Predicted: dog") and overlaid directly on the waveform plot for clear visual interpretation. The waveform represents the amplitude variations of the input audio signal over time, enabling users to correlate the acoustic pattern with the predicted class. This visualization demonstrates the system's capability to perform remote inference and deliver instant, interpretable prediction results.

Comparative Analysis

Table 1 Comparative analysis is essential for understanding the performance, reliability, and effectiveness of different machine learning and deep learning models used in sound signal classification. In this project, multiple classification models including KNN, DTC, LDA, AdaBoost Classifier, and the proposed CNN–RNN hybrid model are trained and evaluated on the same multi-feature dataset extracted using MFCCs, Chroma, Mel-Spectrogram, Tonnetz, Spectral Contrast, and other spectral–temporal features. Each model is assessed using standard performance metrics such as Accuracy, Precision, Recall, and F1-Score, and their behavior is further visualized through confusion matrices and ROC curves. From the results, traditional models such as KNN and DTC demonstrate moderate performance and are able to capture basic patterns in the audio features. However, these models struggle with classes that exhibit overlapping frequency characteristics particularly commands such as bird, cat, dog, down, and left. LDA, being a linear model, performs the weakest because it fails to model complex non-linear acoustic variations. AdaBoost improves performance through its boosting mechanism but still lacks robustness against noisy and high-dimensional audio data. In contrast, the proposed CNN–RNN hybrid model significantly outperforms all baseline approaches. The CNN layers efficiently extract spatial–frequency patterns from spectrograms, while LSTM layers capture sequential and temporal dependencies in sound waves. This combination allows the model to learn both short-term and long-term acoustic structures, resulting in almost perfect classification accuracy. The CNN–RNN model achieves an AUC of 1.00 for all classes, and its confusion matrix shows near-complete dominance along the diagonal, indicating highly accurate predictions and minimal misclassification.

The superior performance of the proposed deep learning architecture highlights the advantage of integrating convolutional and recurrent components to handle complex, non-linear, and time-dependent audio features.

Table 1: Comparative analysis of classification models for sound signal classification

Model	Accuracy	Precision	Recall	F1-Score
KNN Model	0.72	0.69	0.72	0.68
DTC Model	0.70	0.67	0.70	0.66
LDA	0.65	0.62	0.65	0.61
AdaBoost Classifier	0.75	0.73	0.75	0.72
Proposed CNN–RNN	**0.99**	**0.99**	**0.99**	**0.99**

5. Conclusion

The proposed Sound Signal Classification System successfully integrates traditional machine learning techniques and a hybrid deep learning architecture to achieve efficient and accurate classification of audio signals. By using a comprehensive feature extraction pipeline with MFCC, Chroma, Mel-Spectrogram, Tonnetz, and other spectral

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

descriptors, the system ensures robust representation of acoustic information. The hybrid CNN–RNN model further enhances performance by learning both frequency-based patterns through convolutional layers and temporal dependencies through LSTM layers. The implementation of a user-friendly Tkinter GUI enabled seamless interaction for dataset upload, preprocessing, EDA visualization, model training, and evaluation. The inclusion of Redis-based authentication ensures secure access control, while the Flask API provides real-time prediction capabilities. Traditional ML models such as KNN, DTC, AdaBoost, and LDA were implemented and evaluated to benchmark performance against the CNN-RNN model, confirming that the proposed deep learning approach achieved superior accuracy and generalization. Overall, the research demonstrates that integrating advanced audio feature extraction with hybrid deep learning architectures can significantly improve the reliability and precision of sound classification systems. The system is modular, scalable, and capable of being extended to various real-world applications where sound-based recognition is essential.

REFERENCES

[1] Begazo, R.; Aguilera, A.; Dongo, I.; Cardinale, Y. A Combined CNN Architecture for Speech Emotion Recognition. Sensors **2024**, 24, 5797. https://doi.org/10.3390/s24175797

[2] Ullah, R.; Wuttisittikulkij, L.; Chaudhary, S.; Parnianifard, A.; Shah, S.; Ibrar, M.; Wahab, F.-E. End-to-End Deep Convolutional Recurrent Models for Noise Robust Waveform Speech Enhancement. Sensors **2022**, 22, 7782. https://doi.org/10.3390/s22207782

[3] Kubanek, M.; Bobulski, J.; Kulawik, J. A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network. Symmetry **2019**, 11, 1185. https://doi.org/10.3390/sym11091185

[4] Pervaiz, A.; Hussain, F.; Israr, H.; Tahir, M.A.; Raja, F.R.; Baloch, N.K.; Ishmanov, F.; Zikria, Y.B. Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data. Sensors **2020**, 20, 2326. https://doi.org/10.3390/s20082326

[5] Vujičić, D.; Damnjanović, Đ.; Marković, D.; Stamenković, Z. Deep Learning System for Speech Command Recognition. Electronics **2025**, 14, 3793. https://doi.org/10.3390/electronics14193793

[6] Torrie, S.; Wright, K.; Lee, D.-J. MultiAVSR: Robust Speech Recognition via Supervised Multi-Task Audio–Visual Learning. Electronics **2025**, 14, 2310. https://doi.org/10.3390/electronics14122310

[7] Lin, Y.-Y.; Zheng, W.-Z.; Chu, W.C.; Han, J.-Y.; Hung, Y.-H.; Ho, G.-M.; Chang, C.-Y.; Lai, Y.-H. A Speech Command Control-Based Recognition System for Dysarthric Patients Based on Deep Learning Technology. Appl. Sci. **2021**, 11, 2477. https://doi.org/10.3390/app11062477

[8] Hussein, H.H.; Karan, O.; Kurnaz, S. Enhancing Driving Control via Speech Recognition Utilizing Influential Parameters in Deep Learning Techniques. Electronics **2025**, 14, 496. https://doi.org/10.3390/electronics14030496

[9] Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. Symmetry **2019**, 11,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

1018. https://doi.org/10.3390/sym11081018.

[10] Ouali, S.; El Garouani, S. Efficient and Robust Arabic Automotive Speech Command Recognition System. Algorithms **2024**, 17, 385. https://doi.org/10.3390/a17090385

[11] Gupta, S.; Mamodiya, U.; Al-Gburi, A.J.A. Speech Recognition-Based Wireless Control System for Mobile Robotics: Design, Implementation, and Analysis. Automation **2025**, 6, 25. https://doi.org/10.3390/automation6030025

[12] Jeon, S.; Kim, M.S. Noise-Robust Multimodal Audio-Visual Speech Recognition System for Speech-Based Interaction Applications. Sensors **2022**, 22, 7738. https://doi.org/10.3390/s22207738

[13] Yang, W.; Li, P.; Yang, W.; Liu, Y.; He, Y.; Petrosian, O.; Davydenko, A. Research on Robust Audio-Visual Speech Recognition Algorithms. Mathematics **2023**, 11, 1733. https://doi.org/10.3390/math11071733

[14] Tamazin, M.; Gouda, A.; Khedr, M. Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients. Appl. Sci. **2019**, 9, 2166. https://doi.org/10.3390/app9102166