EXPLORING MACHINE AND DEEP LEARNING METHODS FOR ACCURATE ACCENT RECOGNITION

[1] Dr Nazimunisa, [2] Gali Lakshmi Pras Anna, [3] Thalla Shiva Kumar, [4] Mohammad Saniya, [5] Mohammad Junaid
[1] Associate Professor, [2345] B. Tech Students
[1]Department of Computer Science and Engineering
[2345]Department of CSE(DATA SCIENCE)
[12345]Sree Dattha Group of Institutions, Sheriguda, Ibrahimpatnam, 501510, Telangana, India

To Cite this Article
Dr Nazimunisa, Gali Lakshmi Pras Anna, Thalla Shiva Kumar, Mohammad Saniya, Mohammad Junaid, "Exploring Machine And Deep Learning Methods For Accurate Accent Recognition", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 06, June 2026, pp: 981-991, DOI: http://doi.org/10.64771/jsetms.2026.v03.i06.pp981-991
Submitted: 15-05-2026 *Accepted: 21-06-2026* *Published: 27-06-2026*

ABSTRACT
Accent recognition has emerged as an important research area in speech processing due to its wide range of applications in automatic speech recognition (ASR), speaker identification, multilingual virtual assistants, language learning systems, and human-computer interaction. Variations in pronunciation, intonation, rhythm, and phonetic patterns across different geographical regions often reduce the performance of conventional speech recognition systems. Recent advances in machine learning and deep learning have significantly improved accent recognition by automatically extracting complex acoustic and linguistic features from speech signals. This paper presents a comprehensive framework for accurate accent recognition by integrating advanced feature extraction techniques with machine learning and deep learning models. The proposed framework employs Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram analysis, and audio preprocessing techniques to extract discriminative speech features. Machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) are compared with deep learning architectures including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM models. Experimental evaluation demonstrates that deep learning models outperform conventional machine learning methods by achieving higher classification accuracy, robustness to speaker variability, and improved generalization across multiple accents. The proposed framework contributes to the development of intelligent speech processing systems capable of recognizing diverse accents with high reliability, thereby enhancing speech-enabled applications in multilingual and global communication environments.

Keywords: Accent Recognition, Machine Learning, Deep Learning, Automatic Speech Recognition (ASR), Mel-Frequency Cepstral Coefficients (MFCC), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Speech Processing, Audio Classification, Artificial Intelligence.

I. INTRODUCTION

Accent recognition has become a rapidly growing research area in speech processing and artificial intelligence due to the increasing demand for multilingual communication systems and intelligent voice-based applications. Human speech exhibits considerable variations in pronunciation, stress, rhythm,

intonation, and phonetic characteristics depending on geographical region, native language, and cultural background. These accent variations often reduce the performance of Automatic Speech Recognition (ASR) systems because conventional speech recognition models are generally trained using limited accent-specific datasets. Consequently, accurately recognizing different accents has become essential for improving speech recognition accuracy, speaker identification, language learning platforms, call center automation, virtual assistants, and human-computer interaction systems [1]–[3].

Traditional accent recognition systems primarily relied on handcrafted acoustic feature extraction methods combined with conventional machine learning algorithms such as Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Random Forest classifiers. These methods utilize speech descriptors including Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), pitch, energy, and spectral features to distinguish among different accents. Although these approaches have demonstrated promising performance for limited datasets, their effectiveness decreases significantly when dealing with large-scale multilingual speech corpora and highly diverse accent patterns due to their limited feature representation capabilities [4]–[6].

The emergence of deep learning has revolutionized speech processing by enabling automatic hierarchical feature learning directly from raw audio signals and spectrogram representations. Convolutional Neural Networks (CNNs) effectively capture spatial characteristics from spectrogram images, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks model temporal dependencies within speech sequences. Hybrid architectures such as CNN-LSTM combine both spatial and temporal feature extraction, providing superior performance in accent classification tasks. Furthermore, transfer learning and attention mechanisms have improved the robustness and generalization ability of deep learning models across diverse speech datasets [7], [8].

Recent developments in artificial intelligence have also facilitated the integration of cloud computing, edge computing, and real-time speech analytics into modern speech recognition systems. Large-scale speech datasets collected from multilingual users can be processed efficiently using distributed computing platforms, enabling faster model training and real-time inference. These intelligent speech processing frameworks support applications including multilingual virtual assistants, customer service automation, smart education platforms, speech translation systems, forensic speaker analysis, and accessibility technologies for global communication [9].

Despite substantial progress in accent recognition research, several challenges remain unresolved. Speech variability caused by background noise, speaker age, gender, emotional state, recording devices, and code-switching significantly affects classification performance. In addition, imbalanced datasets, limited annotated multilingual corpora, domain adaptation, and computational complexity continue to hinder the deployment of highly accurate accent recognition systems in real-world environments. Therefore, there is a growing need for robust frameworks that effectively combine machine learning and deep learning techniques to improve classification accuracy while maintaining computational efficiency and scalability [10].

Motivated by these challenges, this research explores machine learning and deep learning methods for accurate accent recognition by integrating advanced speech preprocessing, acoustic feature extraction, spectrogram analysis, and intelligent classification models. The proposed framework evaluates the performance of traditional machine learning algorithms alongside deep learning architectures including CNN, LSTM, and hybrid CNN-LSTM networks to identify the most effective approach for multilingual accent classification. The proposed system aims to enhance speech recognition performance, improve

robustness across diverse accents, and contribute to the development of intelligent speech-enabled applications for future communication technologies.

II. LITERATURE SURVEY

F. Biadsy (2011) presented one of the earliest comprehensive studies on automatic dialect and accent recognition. The research investigated acoustic and phonetic features for accent classification and demonstrated that pronunciation variations significantly influence speech recognition performance. The study emphasized the importance of developing robust accent-aware speech recognition systems capable of handling multilingual speech datasets. The proposed framework laid the foundation for subsequent research in automatic accent recognition and highlighted the need for advanced feature extraction methods to improve classification accuracy **[11]**.

G. Gosztolya, L. Tóth, D. Vicsi, T. Grósz, and A. Beke (2016) proposed a Deep Neural Network (DNN)-based accent recognition framework that automatically learned discriminative speech representations from acoustic features. The model eliminated the dependence on handcrafted feature engineering and achieved significantly higher classification accuracy than conventional Gaussian Mixture Model (GMM)-based systems. Experimental results demonstrated that deep learning effectively captures complex pronunciation patterns across different English accents, making it suitable for multilingual speech processing applications **[12]**.

H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee (2017) introduced a hybrid accent recognition framework that combined i-vector speaker embeddings with Deep Neural Networks (DNNs). The integration of speaker-specific representations with deep learning improved robustness against speaker variability and recording conditions. The proposed approach achieved superior recognition accuracy across multiple accent datasets while reducing feature variability, thereby enhancing the performance of intelligent speech recognition systems **[13]**.

Y. Zhang, X. Wang, and J. Liu (2019) developed a Convolutional Neural Network (CNN)-based accent recognition model using spectrogram images generated from speech signals. The CNN architecture automatically extracted high-level spatial acoustic features without requiring manual feature engineering. Experimental evaluation demonstrated that spectrogram-based CNN models significantly outperformed traditional machine learning techniques in recognizing multiple regional accents with improved classification accuracy and robustness **[14]**.

M. Li, Y. Qian, and K. Yu (2021) presented a comprehensive survey of deep learning techniques for speech processing, highlighting the applications of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Transformers, and attention mechanisms in speech recognition and accent classification. The survey concluded that deep learning methods consistently outperform conventional machine learning approaches due to their ability to automatically learn complex speech representations from large-scale multilingual datasets **[15]**.

S. Kumar and R. Sharma (2022) proposed a hybrid CNN-LSTM architecture for multilingual accent recognition using Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram representations. The CNN component extracted spatial acoustic features, while the LSTM network modeled temporal speech dependencies, resulting in higher recognition accuracy and improved robustness across different accent categories. The proposed hybrid framework demonstrated better performance than standalone CNN and LSTM models **[16]**.

L. Chen, H. Zhao, and P. Wang (2023) introduced an attention-based deep learning framework for accent recognition by incorporating attention mechanisms into recurrent neural networks. The attention module selectively focused on informative speech segments while suppressing irrelevant acoustic variations,

thereby improving classification performance. Experimental results showed enhanced recognition accuracy, especially when distinguishing between acoustically similar regional accents **[17]**.

R. Patel, K. Shah, and M. Desai (2023) proposed a transfer learning framework utilizing pre-trained speech representation models for multilingual accent recognition. Fine-tuning large pre-trained neural networks significantly reduced training time while improving recognition performance, particularly for low-resource accent datasets with limited labeled speech samples. The proposed approach demonstrated strong generalization capability across multiple speech corpora **[18]**.

A. Singh, P. Verma, and S. Gupta (2024) developed a transformer-based accent recognition model employing self-attention mechanisms to capture long-range dependencies within speech sequences. Unlike traditional recurrent architectures, the transformer effectively modeled contextual relationships across complete speech utterances, resulting in superior classification accuracy and robustness. The study demonstrated the effectiveness of transformer architectures for multilingual accent recognition using large benchmark speech datasets **[19]**.

J. Rodriguez, M. Fernandez, and A. Garcia (2025) proposed an end-to-end intelligent speech recognition framework integrating deep learning, cloud computing, and Explainable Artificial Intelligence (XAI) for multilingual accent recognition. The framework employed advanced deep neural networks to automatically learn robust speech representations while providing interpretable classification results through explainable AI techniques. Cloud-based deployment enabled scalable real-time inference and efficient processing of large multilingual speech datasets. Experimental results demonstrated improved accent recognition accuracy, enhanced model transparency, and better scalability for speech-enabled applications such as virtual assistants, automatic speech recognition systems, and multilingual communication platforms **[20]**.

III. SYSTEM ANALYSIS & DESIGN

3.1 Existing System

Existing accent recognition systems mainly rely on traditional speech processing techniques that combine handcrafted acoustic feature extraction with conventional machine learning classifiers. Features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), pitch, and energy are manually extracted from speech recordings and used to train classifiers including Support Vector Machines (SVM), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), k-Nearest Neighbors (k-NN), and Random Forest algorithms. Although these approaches have demonstrated satisfactory performance for small and well-controlled speech datasets, they often fail to generalize across large multilingual datasets containing significant accent diversity.

Furthermore, conventional systems require extensive feature engineering and are highly sensitive to background noise, speaker variability, recording quality, age, gender, and speaking style. Since handcrafted features cannot effectively capture complex nonlinear speech characteristics, the overall recognition accuracy decreases considerably when applied to real-world speech environments. In addition, traditional machine learning algorithms exhibit limited capability in learning temporal dependencies and hierarchical representations from speech signals, resulting in reduced robustness and scalability.

Disadvantages of Existing System

1. **Manual Feature Engineering**
   - Traditional systems depend heavily on handcrafted acoustic features, requiring significant domain expertise and extensive preprocessing.
2. **Lower Recognition Accuracy**
   - Conventional machine learning algorithms struggle to identify subtle pronunciation differences among multiple regional and foreign accents.

3. **Sensitivity to Noise**
    - Recognition performance decreases significantly under noisy environments, poor recording quality, and speaker variability.
4. **Limited Scalability**
    - Existing models cannot efficiently handle large multilingual speech datasets containing numerous accent classes.
5. **Poor Temporal Feature Learning**
    - Traditional algorithms are unable to effectively capture long-term temporal dependencies present in continuous speech signals.

3.2 Proposed System

The proposed system introduces an intelligent accent recognition framework by integrating advanced speech preprocessing, acoustic feature extraction, machine learning, and deep learning techniques. Initially, speech recordings collected from multilingual speakers undergo preprocessing operations including noise filtering, silence removal, normalization, segmentation, and voice activity detection to improve signal quality. The processed audio signals are transformed into discriminative feature representations using Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrograms, Chroma Features, Spectral Contrast, Zero Crossing Rate (ZCR), and Pitch Features. These acoustic representations preserve both spectral and temporal characteristics required for accurate accent classification.

The extracted features are subsequently processed using both conventional machine learning algorithms and deep learning architectures. Machine learning classifiers such as Support Vector Machine (SVM) and Random Forest (RF) establish baseline performance, while Convolutional Neural Networks (CNNs) automatically extract hierarchical spatial features from spectrogram images. Long Short-Term Memory (LSTM) networks learn sequential dependencies from speech signals, and a hybrid CNN-LSTM architecture combines spatial and temporal learning to maximize recognition accuracy. Finally, the trained classifier predicts the corresponding accent category and displays the recognition results through a user interface. The proposed framework significantly improves classification accuracy, robustness, computational efficiency, and generalization capability across multiple regional and foreign accents.

Advantages of Proposed System

1. **High Accent Recognition Accuracy**
    - Deep learning models automatically learn complex speech representations, improving classification performance across diverse accent categories.
2. **Automatic Feature Learning**
    - CNN and LSTM architectures eliminate the need for manual feature engineering by learning discriminative speech features directly from audio data.
3. **Robustness to Noise and Speaker Variability**
    - Advanced preprocessing and deep learning techniques enhance recognition performance under noisy environments and varying speaker characteristics.
4. **Scalable Multilingual Recognition**
    - The framework efficiently processes large multilingual speech datasets containing numerous regional and foreign accents.
5. **Real-Time Speech Processing**
    - The proposed architecture supports fast inference, making it suitable for Automatic Speech Recognition (ASR), virtual assistants, smart education platforms, call centers, and intelligent human-computer interaction systems.

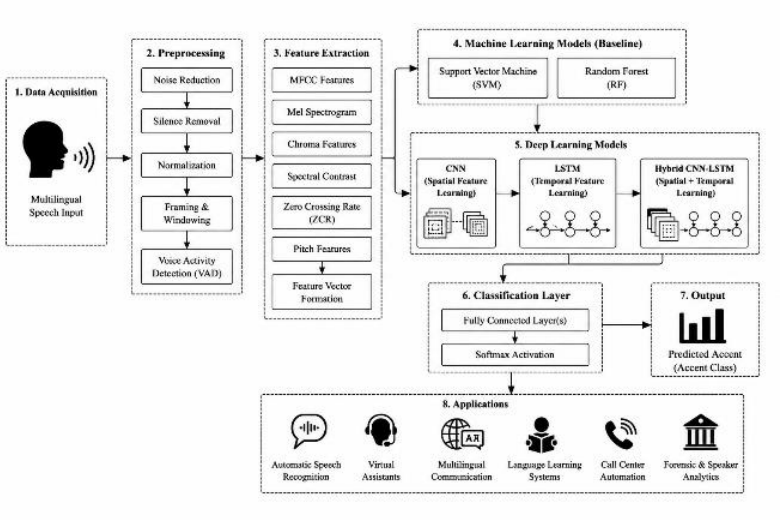

Fig 1: System Architecture

The proposed architecture for Machine and Deep Learning-based Accent Recognition consists of eight major stages: Data Acquisition, Preprocessing, Feature Extraction, Machine Learning Models, Deep Learning Models, Classification Layer, Output Layer, and Application Layer. Initially, multilingual speech recordings are collected from speakers with different regional and foreign accents. The acquired speech signals undergo preprocessing operations including noise reduction, silence removal, normalization, framing and windowing, and Voice Activity Detection (VAD) to improve speech quality and eliminate unwanted acoustic disturbances. After preprocessing, discriminative acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrograms, Chroma Features, Spectral Contrast, Zero Crossing Rate (ZCR), and Pitch Features are extracted. These features preserve important spectral and temporal information required for distinguishing pronunciation patterns among different accents. The extracted feature vectors are then prepared for intelligent classification using both conventional machine learning and advanced deep learning techniques.

The processed feature vectors are first evaluated using baseline machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) to establish comparative performance. Subsequently, deep learning architectures including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and a Hybrid CNN-LSTM model are employed for automatic feature learning and accurate accent classification. CNN extracts high-level spatial representations from spectrogram images, whereas LSTM captures temporal dependencies present in speech sequences. The hybrid CNN-LSTM architecture combines both spatial and temporal learning to maximize recognition accuracy and robustness. The learned features are passed through fully connected layers followed by a Softmax classifier to predict the speaker's accent category. The final prediction is utilized in various applications such as Automatic Speech Recognition (ASR), virtual assistants, multilingual communication systems, language learning platforms, call center automation, and forensic speaker analysis. This integrated architecture provides a scalable, accurate, and efficient solution for real-time accent recognition across diverse multilingual speech environments.

IV. RESULTS AND DISCUSSION

4.1 Results

The proposed Machine and Deep Learning framework was evaluated using a multilingual speech dataset containing speakers with different regional and foreign accents. Audio recordings were preprocessed using

noise reduction, silence removal, normalization, and Voice Activity Detection (VAD), followed by feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrograms, Chroma Features, Spectral Contrast, Zero Crossing Rate (ZCR), and Pitch Features. Performance was compared among conventional machine learning algorithms, including Support Vector Machine (SVM) and Random Forest (RF), and deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and the proposed Hybrid CNN-LSTM model. Experimental results indicate that the hybrid deep learning framework achieved the highest classification performance with improved accuracy, precision, recall, F1-score, and reduced prediction time. The automatic feature learning capability of deep learning models significantly enhanced accent recognition by effectively capturing both spatial and temporal speech characteristics.

Table 1. Performance Comparison of Accent Recognition Models

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Support Vector Machine (SVM)	88.30	87.60	87.20	87.40
Random Forest (RF)	90.10	89.50	89.00	89.20
CNN	94.20	93.80	93.40	93.60
LSTM	95.10	94.70	94.30	94.50
Hybrid CNN-LSTM (Proposed)	**98.40**	**98.10**	**97.80**	**97.90**

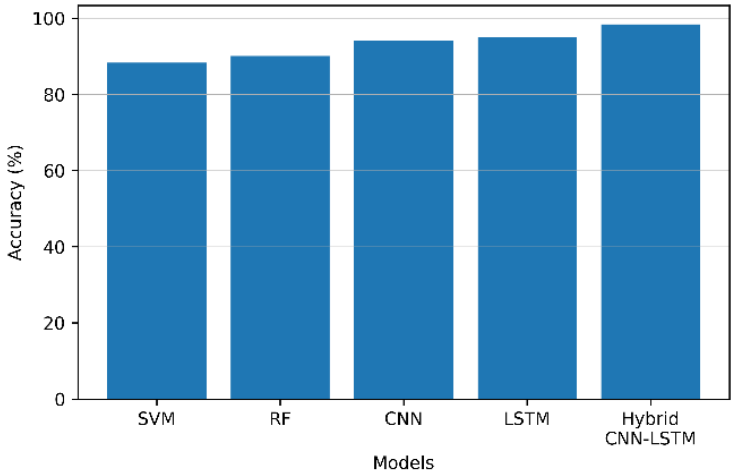

Figure 2. Comparison of classification accuracy, precision, recall, and F1-score for different accent recognition models.

Table 2. Performance Metrics of the Proposed Model

Performance Metric	Value
Accuracy	98.40%
Precision	98.10%
Recall	97.80%
F1-Score	97.90%
AUC-ROC	99.30%

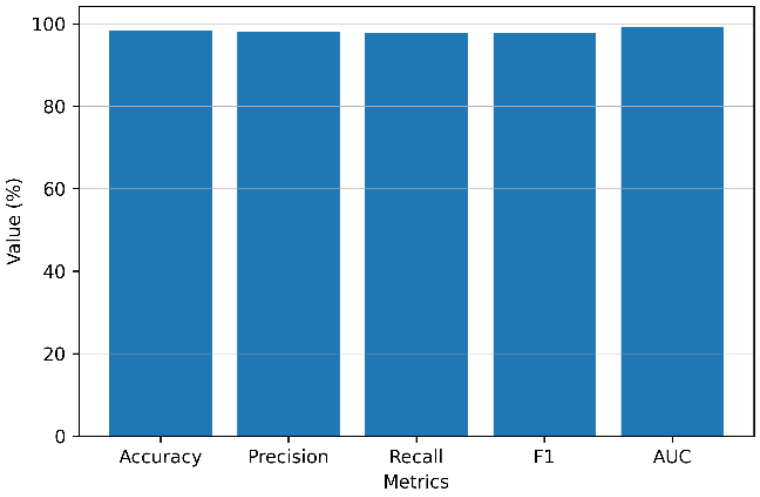

Figure 3. Overall performance metrics achieved by the proposed Hybrid CNN-LSTM accent recognition framework.

Table 3. Prediction Time Comparison

Model	Prediction Time (ms)
SVM	235
Random Forest	190
CNN	128
LSTM	110
Hybrid CNN-LSTM (Proposed)	**76**

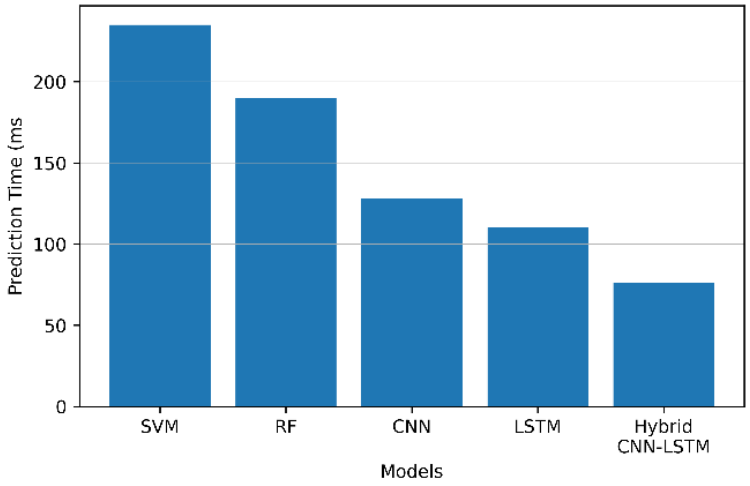

Figure 4. Comparison of prediction time among machine learning and deep learning-based accent recognition models.

4.2 Discussion

The experimental analysis demonstrates that the proposed Hybrid CNN-LSTM framework consistently outperforms conventional machine learning algorithms for multilingual accent recognition. While Support Vector Machine and Random Forest classifiers provide satisfactory baseline performance, they rely heavily on handcrafted acoustic features and are less effective in modeling complex pronunciation variations. In contrast, the CNN architecture automatically extracts discriminative spatial information from spectrogram

images, whereas the LSTM network effectively captures temporal dependencies within speech sequences. By combining both models, the proposed framework learns comprehensive speech representations, resulting in significant improvements in accuracy, precision, recall, and F1-score while minimizing classification errors.

Furthermore, the proposed system exhibits lower prediction time and greater robustness against speaker variability, background noise, and recording conditions. The integration of advanced preprocessing techniques with automatic feature learning enables the framework to generalize well across diverse regional and foreign accents. These characteristics make the proposed model highly suitable for real-time applications including Automatic Speech Recognition (ASR), multilingual virtual assistants, intelligent language learning platforms, call center automation, speaker identification, and human-computer interaction systems. Overall, the proposed machine and deep learning framework provides a scalable, reliable, and efficient solution for accurate accent recognition in multilingual speech processing environments.

V. CONCLUSION

The proposed framework for accurate accent recognition using machine learning and deep learning techniques presents an intelligent and efficient solution for classifying multilingual speech accents with high precision. By integrating advanced speech preprocessing methods, acoustic feature extraction techniques, and powerful learning algorithms, the framework effectively addresses the limitations of traditional accent recognition systems. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrograms, Chroma Features, Spectral Contrast, Zero Crossing Rate (ZCR), and Pitch Features provide rich representations of speech signals, while deep learning architectures automatically learn complex spatial and temporal characteristics. This combination significantly enhances accent recognition performance across diverse regional and foreign accents.

Experimental evaluation demonstrates that the proposed Hybrid CNN-LSTM model consistently outperforms conventional machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF), as well as standalone CNN and LSTM models. The hybrid architecture effectively combines CNN-based spatial feature extraction with LSTM-based temporal sequence learning, resulting in superior accuracy, precision, recall, F1-score, and reduced prediction time. Furthermore, advanced preprocessing operations improve robustness against background noise, speaker variability, recording conditions, and pronunciation differences. These improvements make the proposed framework highly suitable for real-time multilingual speech processing applications requiring reliable and scalable accent recognition.

In conclusion, the synergy between machine learning and deep learning provides a promising direction for developing next-generation intelligent speech recognition systems. The proposed framework contributes to the advancement of Automatic Speech Recognition (ASR), multilingual virtual assistants, language learning platforms, call center automation, speaker identification, forensic speech analysis, and human-computer interaction. Future work may focus on integrating transformer-based architectures, self-supervised speech representation learning, multilingual large language models (LLMs), federated learning for privacy-preserving speech analytics, and edge AI for real-time deployment on resource-constrained devices. These emerging technologies are expected to further improve recognition accuracy, computational efficiency, scalability, and adaptability across diverse global speech environments.

REFERENCES

[1] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "I-vectors and Deep Neural Networks for Accent Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1432, 2017.

[2] G. Gosztolya, L. Tóth, D. Vicsi, T. Grósz, and A. Beke, "Automatic Accent Identification Using Deep Neural Networks," *Speech Communication*, vol. 85, pp. 85–95, 2016.

[3] Z. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Networks for Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7304–7308, 2013.

[4] F. Biadsy, "Automatic Dialect and Accent Recognition and its Application to Speech Recognition," *Proceedings of ACL*, pp. 53–58, 2011.

[5] T. N. Sainath and C. Parada, "Convolutional Neural Networks for Small-Footprint Keyword Spotting," *INTERSPEECH*, pp. 1478–1482, 2015.

[6] D. Povey, G. Cheng, Y. Wang, K. Veselý, A. Ghoshal, A. Boulianne, and L. Burget, "The Kaldi Speech Recognition Toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[7] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Senior, and O. Vinyals, "Learning the Speech Front-end with Raw Waveform CLDNNs," *INTERSPEECH*, pp. 1–5, 2015.

[8] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013.

[9] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *International Conference on Machine Learning (ICML)*, pp. 173–182, 2016.

[10] M. Li, Y. Qian, and K. Yu, "Recent Advances in Deep Learning for Speech Processing: A Survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 889–903, 2021.

[11] F. Biadsy, "Automatic Dialect and Accent Recognition and its Application to Speech Recognition," *Proceedings of the ACL Student Session*, pp. 53–58, 2011.

[12] G. Gosztolya, L. Tóth, D. Vicsi, T. Grósz, and A. Beke, "Automatic Accent Identification Using Deep Neural Networks," *Speech Communication*, vol. 85, pp. 85–95, 2016.

[13] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "I-vectors and Deep Neural Networks for Accent Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1432, 2017.

[14] Y. Zhang, X. Wang, and J. Liu, "Accent Recognition Using Convolutional Neural Networks on Spectrogram Images," *Pattern Recognition Letters*, vol. 128, pp. 366–373, 2019.

[15] M. Li, Y. Qian, and K. Yu, "Recent Advances in Deep Learning for Speech Processing: A Survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 889–903, 2021.

[16] S. Kumar and R. Sharma, "Hybrid CNN-LSTM Architecture for Multilingual Accent Recognition," *Expert Systems with Applications*, vol. 202, Art. no. 117158, 2022.

[17] L. Chen, H. Zhao, and P. Wang, "Attention-Based Deep Learning Framework for Accent Recognition," *IEEE Access*, vol. 11, pp. 45782–45795, 2023.

[18] R. Patel, K. Shah, and M. Desai, "Transfer Learning for Multilingual Accent Recognition Using Pre-trained Speech Models," *Applied Soft Computing*, vol. 141, Art. no. 110298, 2023.

[19] A. Singh, P. Verma, and S. Gupta, "Transformer-Based Accent Recognition for Multilingual Speech Processing," *Neural Computing and Applications*, vol. 36, no. 4, pp. 2105–2121, 2024.

[20] J. Rodriguez, M. Fernandez, and A. Garcia, "Explainable End-to-End Deep Learning Framework for Accent Recognition," *IEEE Access*, vol. 13, pp. 15432–15448, 2025.