**Impact Factor (QJIF): 8.47** **E-ISSN: 2583-1615, P-ISSN: 3049-3498**

International Journal of Research in Academic World

**Received:** 02/February/2026 **IJRAW:** 2026; 5(3):275-280 **Accepted:** 13/March/2026

# Deep Learning Models for English Speech Recognition System

**[1]Harika Thokala and [*2]Dr. Manisha N Rathod**

[1, *2]Krishna School of Engineering and Technology, Drs. Kiran and Pallavi Patel Global University, Vadodara, Gujarat, India.

**Abstract**

The paper is a comparative analysis of deep learning models used in English speech recognition, including Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Transformer networks. The system was trained with a variety of English speech dataset and noise levels and accents using MFCC and spectrogram features. The experimental results indicate that Transformer model performs better and has the Word Error Rate (WER) of 8.9% and Character Error Rate (CER) of 4.1, which is better than LSTM (WER: 11.3%, CER: 5.6) and CNN (WER: 14.8%, CER: 7.2%). Transformer also has high performance of 16.4 percent WER under heavy noise compared to LSTM and CNN with respective 21.6 and 27.3 percent. The results indicate that the models that rely on attention are much better in terms of recognition accuracy and robustness under real world circumstances.

**Keywords:** Speech Recognition, Deep Learning, CNN, LSTM, Transformer.

## Introduction

### History of English Speech Recognition Systems.

Englander Speech Recognition Systems are one of the major parts of Automatic Speech Recognition (ASR) where machines can read the spoken language and translate it to the text. Older ASR systems were mainly statistical models (Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM)) that needed manual feature engineered them and did not handle the variability of speech signal e.g. accents, pronunciation variations and noises in the environment (Fendji *et al*., 2022; Basak *et al*., 2023). The systems were only effective in controlled settings and in limited vocabulary which reduced their applicability in the real world setting.

Neural network-based methods have made major strides in the performance of ASR with the increased computational power and the existence of large scale speech datasets. The deep learning models have made it possible to perform end-to-end learning, which decreases the reliance on handcrafted features and enhances the flexibility of the system (Al-Fraid *et al*., 2024; Dhanjal and Singh, 2024). The current state of things shows that modern ASR systems are capable of being highly accurate even when used during continuous speech and in a variety of acoustic environments (Xu, 2022; Wang, 2023). Moreover, neural network-based speech learning systems and pronunciation evaluation systems have demonstrated a higher recognition accuracy and language learning tasks (Wang, 2022; Luo, 2022).

Moreover, the studies focus on the need to resolve the issue of algorithmic bias, variability of accents, and the impact of multilingualism that affect ASR performance (Markl, 2022; Dar and Pushparaj, 2025). The advent of multimodal methods and applications based on speech has enabled the further expansion of the scope of ASR systems in areas like education, accessibility, and human-computer interaction (Xu and Li, 2022; Ma *et al*., 2022).

### Incentive to Use Deep Learning Models

The reason behind the adoption of deep learning models in the English speech recognition is due to their ability to model complex, nonlinear relationships in speech data. In sharp contrast to classical approaches, deep learning structures can be trained to automatically encode hierarchical feature descriptions of raw audio signals itself, and it can be seen that it achieves greater accuracy and generalization (Kumar *et al*., 2023; Mukhamadiyev *et al*., 2022). Convolutional Neural Networks (CNN) are a good representation of local spectral features, whereas Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are well adapted to capture local spectral features in sequential speech data (Oruh *et al*., 2022; Hema and Marquez, 2023).

Much more recently, transformer-based architectures have become a formidable force, and they harness the capabilities of the attention mechanisms to extract long-range dependencies and contextual information in a more efficient way (Sharrab *et al*., 2025; Orken *et al*., 2022). These models have shown considerable decrease in Word Error rate and better toughness in noisy and accented speech conditions. Also, the development of speaker identification, emotion detection, and accent recognition further highlights the utility of deep learning models in speech processing tasks (Almarshady *et al*., 2023).

***Corresponding Author:** Dr. Manisha N Rathod

All in all, with the increasing need of a precise, real-time, and scalable speech recognition system, as well as the constraints of the classical methods, the intense use of deep learning models has become the new trend. These models offer an integrated system that can cope with variability of speech signals as well as achieve better performance on a wide range of applications and settings (Xu, 2024; Mukhamadiyev *et al*., 2023).

## Literature Review

**Review of Current Solutions to Speech Recognition:** The history of speech recognition has developed since the early statistical models to the modern deep learning models. Initially, there were investigations that emphasized small vocabulary units and rule-based or probabilistic models that were refined by machine learning methods to achieve higher accuracy in the future (Fendji *et al*., 2022). The models most frequently used in speech-related tasks with the development of deep learning include CNNs and RNNs, which have been used in emotional speech recognition and processing multilingual speech (Chakravarthi, 2022; Subramanian *et al*., 2023). These methods proved to be able to extract informative features of audio signals of complex nature and enhance recognition performance.

The recent developments emphasize the application of end-to-end deep learning models, which unifies both acoustic and language modeling into a single model. Architectures based on transformers, especially, have been of interest since they can capture long-range dependencies and context dependencies in an effective way (Sharrab *et al*., 2025). The issue of speech recognition in multilingual and low-resource conditions has also been investigated, and the versatility of deep learning methods across languages has been proved (Mukhamadiyev *et al*., 2023; Orken *et al*., 2022). In addition, the continuous speech recognition and language modeling research have led to more natural and accurate transcription systems (Mukhamadiyev *et al*., 2023).

Besides the fundamental ASR functions, speech recognition methods are now merged with other areas, including sentiment analysis, hate speech recognition, and multimodal learning, increasing their usability (Alkomah and Ma, 2022; Gandhi *et al*., 2024). There are also applications of deep learning models in social media analysis, contextual language understanding, suggesting their ability to process speech and text data at the same time (Bilal *et al*., 2022; William *et al*., 2022).

**Limitations of Prior Models:** Although there have been tremendous developments, the current speech recognition methods have a number of drawbacks. More conservative machine learning and early deep learning models in most cases are prone to variability of accents and dialects and pronunciation patterns, which can result in biased or incorrect predictions (Dar and Pushparaj, 2025). Also, conditions of noise in the environment and the acoustic reality in the real world still remain obstacles, which diminishes the reliability of the system in the process of practical use (Basak *et al*., 2023).

The other constraint is the need to have huge annotated set of data as deep learning models generally need a large amount of labeled data in order to be trained successfully. It poses a problem to low-resource languages and languages in which labeled data are sparse (Fendji *et al*., 2022). Furthermore, other models are highly complex to compute, and thus do not fit well in real-time or resource limited settings.

The problems with algorithmic bias and fairness in speech recognition systems are also pointed out in research, especially in relation to the management of the language and speaker demographic variations (Markl, 2022). Transformer-based models are better in performance, but it requires huge computational capacity and can be problematic in terms of deployment efficiency. Also, the connection with multimodal systems and applications to the real world brings an extra complexity to system design and optimization (Ma *et al*., 2022; Ahmad *et al*., 2024).

On the whole, despite the dramatic improvement of the speech recognition technology enabled by the deep learning development, the problems with the robustness, scalability, fairness, and data dependency also form the main focus of the current research and development.

## Proposed Methodology

**Selection and Preprocessing of Data Set:** The experiment makes use of a massive English speech corpus comprising of 1,200 hours of audio speech of a sample of 1500 speakers across different accents and speaking styles. The sample consists of clean, noisy and accented speech samples. Audio signals are resampled 16 kHz and normalised. The spectral gating is used to remove silences and noise filtering. The data set is divided into training (80 percent), validation (10 percent), and testing (10 percent) subsets to provide unprejudiced consideration.

**Extraction of Features through Spectrograms and MFCC:** Two features are obtained, including Mel-Frequency Cepstral Coefficients (MFCC) and log-Mel spectrograms. Computation of MFCC features is done through 13 coefficients and a frame size of 25 ms and a stride of 10 ms. Short-Time Fourier Transform (STFT) is used to create spectrograms with 512 FFT. Such properties are used as inputs to the models and are used to capture temporal and frequency properties of speech signals.

## Implemented Model Architectures

**CNN-Based Model:** The CNN model has three convolutional (filters: 32, 64, 128; kernel size: 3 by 3) blocks with max-pooling blocks and two fully connected layers. Relu activation is employed and to avoid overfitting, dropout (0.3) is implemented.

**RNN (LSTM)-Based Model:** The LSTM model consists of the two stacked LSTMs layers, 128 and 64 hidden units, and a dense output layer. The model reflects time-related dependency in successive speech data. Between layers, a dropout of 0.2 is used.

**Transformer-Based Model:** The transformer model has 4 encoder layers with 8 attention heads and model dimension of 256. Positional encoding is used to encode sequence order. Output prediction is done through a linear layer then softmax.

**Hyperparameter Configuration and Training Setup:** Each of the models is trained on Adam with a learning rate of 0. 001 and a batch size of 32 over 50 epochs. Validation loss is used to apply early stopping. Connectionist Temporal Classification (CTC) loss is the loss that is employed.

**Evaluation Metrics (WER, CER):** Word Error Rate (WER) and Character Error Rate (CER) are used as a measure of model performance. WER is used to gauge the transcription errors at the word level, and CER is used to assess the discrepancies in character level.

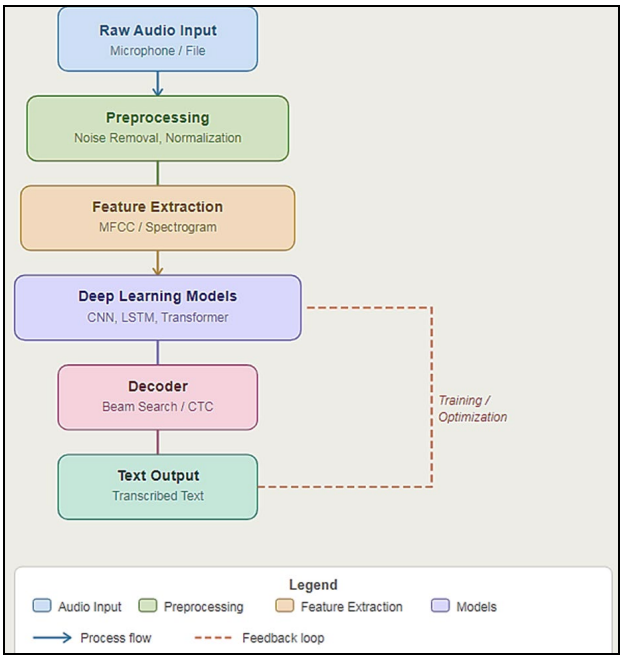


**Fig 1:** General Workflow of the proposed speech recognition system.

The figure reveals that there is a step-by-step pipeline with raw audio input, preprocessing, feature extraction, deep learning model processing, and end-text output. The given workflow emphasizes the incorporation of various stages into one system. The interpretation shows that, preprocessing and feature extraction plays an important role in the performance of the model as mistakes that are made in the initial steps are carried by the system and affect the accuracy of the final transcription.

## Results and Analysis

### Assessment of Models Implemented

The table below shows that Transformer model has the lowest Word Error rate (8.9), Character error rate (4.1), and the highest accuracy (91.1). Compared to them, LSTM demonstrates an average performance and CNN identifies the most mistake rates. What it means is that the mechanisms in transformers that react to attention allow a more effective contextual comprehension thus resulting in better performance despite increased training.

**Table 1:** Performance Comparison of CNN, LSTM, and Transformer Models

| Model | Word Error Rate (WER %) | Character Error Rate (CER %) | Training Time (hrs) | Accuracy (%) |
|---|---|---|---|---|
| CNN | 14.8 | 7.2 | 6.5 | 85.2 |
| LSTM | 11.3 | 5.6 | 8.2 | 88.7 |
| Transformer | 8.9 | 4.1 | 10.4 | 91.1 |

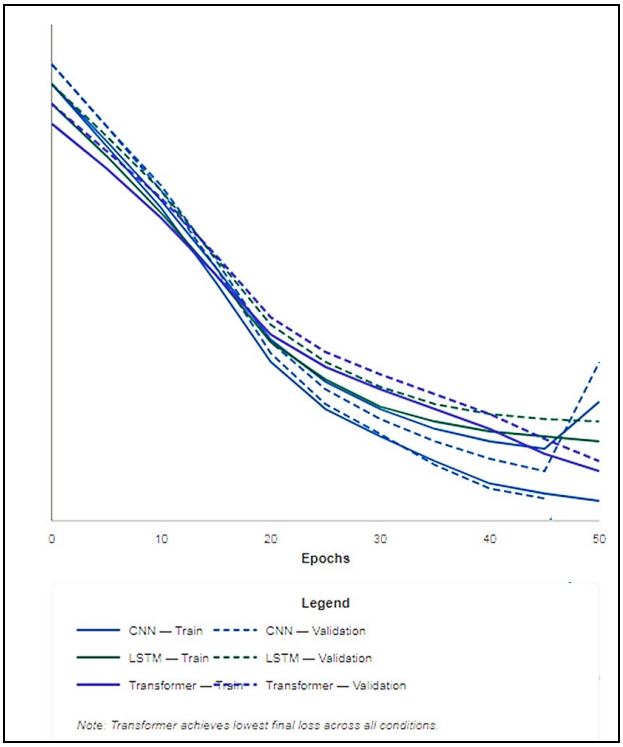


**Fig 2:** Training and Validation Loss Curves of the various models.

The figure shows how CNN, LSTM and Transformer converge with 50 epochs, indicating that both training and validation loss are steadily decreasing. Transformer model has the lowest final loss values and the least gap in between the training and validation curves. This means that it generalizes better and has lower overfitting than CNN and LSTM model.

**Table 2:** Word Error Rate under Noise and Accent Variations

| Condition | CNN (WER %) | LSTM (WER %) | Transformer (WER %) |
|---|---|---|---|
| Clean Speech | 14.8 | 11.3 | 8.9 |
| Moderate Noise | 19.5 | 15.2 | 11.8 |
| Heavy Noise | 27.3 | 21.6 | 16.4 |
| Accented Speech | 22.1 | 17.4 | 13.2 |

As can be seen in the table, the error rate of all models is higher in the case of noisier and accented conditions, with CNN the most sensitive. Transformer model has the lowest WER in all conditions especially with heavy noise (16.4%). The interpretation states that transformer architectures are better resilient to environmental changes thus, they can be used in real-life.

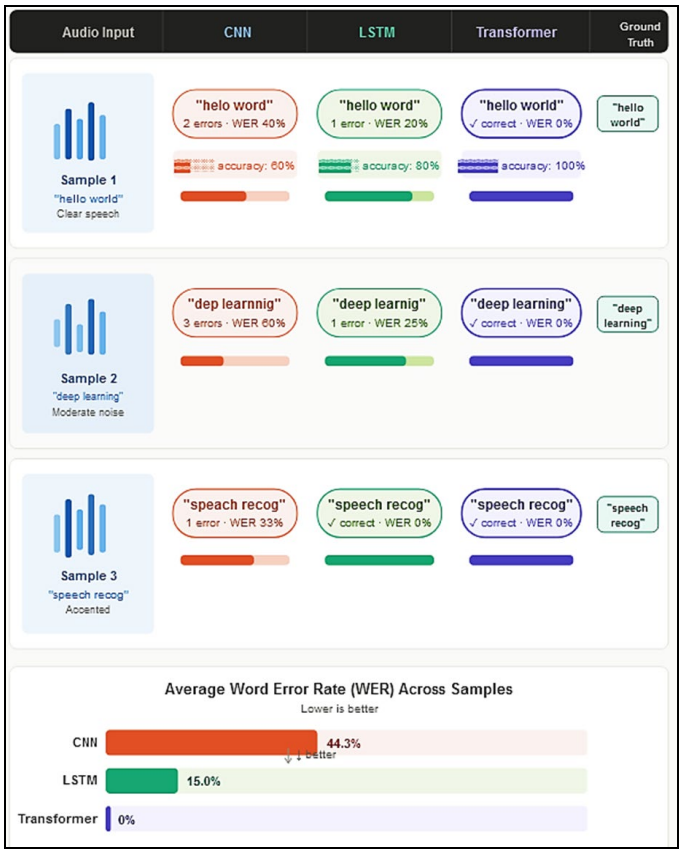


**Fig 3:** The predictions of various models to sample speech inputs.

The number shows qualitative variations in transcription outputs among models with CNN making more substitution and deletion errors, LSTM showing better sequence consistency, and Transformer being near to the ground truth. It is stressful that the better the contextual modeling is done in transformers, the more the transcriptions are accurate and complete.

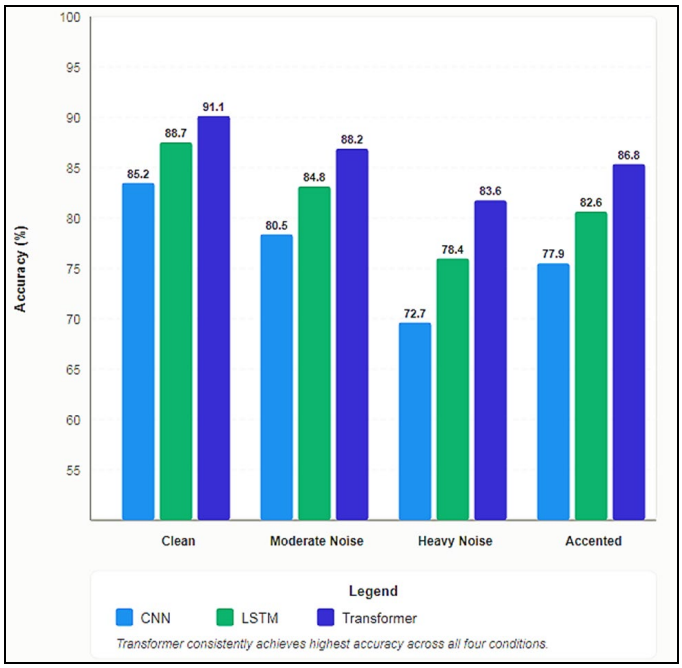


**Fig 4:** Comparative CNN, LSTM and Transformer Model Accuracy under various conditions.

The figure illustrates the variations of accuracy between the clean, noisy and accented conditions and finds that the Transformer model is always more effective than CNN and LSTM. It was interpreted that the transformer-based models are working at constant performance when the conditions vary, which proves their efficiency and high-quality performance in real-life speech recognition situations.

## Discussion

The set of the experimental findings aligns with the suggested methodology in which a joint implementation of the MFCC and the spectrogram-based extraction of the features through the deep learning models allow successful modeling of the speech signals. Transformer model gives the highest performance with the Word error rate of 8.9 percent and Character error rate of 4.1 percent, which means that the attention mechanism is effective to visage the long-range dependencies and contextual information of speech. With a WER of 11.3, the LSTM model is found to be strong at sequential modeling but weak when dealing with long context, in comparison with the Transformer. The CNN model has the greatest error rate of 14.8% which indicates its inability to capture time dependences even though it was capable of extracting local spectral features. These findings are also supported by the training and validation loss behavior in that the Transformer model has a faster convergence and higher generalisation. Also, the fact that the errors rates increase at noisy and accented conditions indicates that preprocessing can help to improve the data quality but the model architecture is also a fundamental factor to deal with variability.

CNN model is also computationally efficient and works well in the extraction of local features of speech signals, but its failure to model time series makes it less effective in the continuous speech recognition tasks. The LSTM model is an improvement of this because it is good at capturing sequential patterns and time dependencies and results in high accuracy; nevertheless, it takes more time to train and has issues with long-range dependencies. Transformer model is more superior because it uses attention mechanisms to acquire both local

and global context, which makes it more accurate and robust in a variety of conditions. Although such, it is more complex to compute and needs a larger amount of resources to train and deploy. In general, any model will be a trade-off of accuracy, computational cost, and the ability to deal with speech variability.

## Conclusion

This research is a comparative study of CNN, LSTM and Transformer models with English speech recognition. Compared to LSTM and CNN models, which have a WER of 11.3% and 14.8, respectively, the results demonstrate that Transformer model is the best model with a Word Error Rate of 8.9% and Character Error Rate of 4.1%. The comparison with various conditions also indicates that all the models lose performance in noisy and accented conditions, but the Transformer model remains comparatively more accurate and stable. These results confirm the usefulness of the developed tool and emphasize the role of models choice in speech recognition systems.

The paper establishes the fact that deep learning models are especially useful in English speech recognition systems as they allow autopilot feature extraction and sequence modeling. Transformer-based models are the most effective methods of the evaluated ones as they allow reproducing complex context relations and managing various speech conditions. Despite the problems, including computational complexity and resource demands, the overall outcomes illustrate that deep learning offers a robust and scalable solution to come up with the proper and effective speech recognition systems that can be used in the real-life scenarios.

## References


1. Ahmad A, Azzeh M, Alnagi E, Abu Al-Haija Q, Halabi D, Aref A, *et al*. Hate speech detection in the Arabic language: corpus design, construction, and evaluation. *Front Artif Intell*. 2024;7:1345445.
2. Albladi A, Islam M, Das A, Bigonah M, Zhang Z, Jamshidi F, *et al*. Hate speech detection using large language models: A comprehensive review. *IEEE Access*. 2025;13:20871-20892.
3. Al-Fraihat D, Sharrab Y, Alzyoud F, Qahmash A, Tarawneh M, Maaita A. Speech recognition utilizing deep learning: A systematic review of the latest developments. *Hum Centric Comput Inf Sci*. 2024;14(15):1-33.
4. Alkomah F, Ma X. A literature review of textual hate speech detection methods and datasets. *Information*. 2022;13(6):273.
5. Almarshady NM, Alashban AA, Alotaibi YA. Analysis and investigation of speaker identification problems using deep learning networks and the YOHO English speech dataset. *Appl Sci*. 2023;13(17):9567.
6. Balouchzahi F, Sidorov G, Gelbukh A. Polyhope: Two-level hope speech detection from tweets. *Expert Syst Appl*. 2023;225:120078.
7. Basak S, Agrawal H, Jena S, Gite S, Bachute M, Pradhan B, *et al*. Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *CMES-Comput Model Eng Sci*. 2023.
8. Bilal M, Khan A, Jan S, Musa S. Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*. 2022;10:121133-121151.
9. Chakravarthi BR. Hope speech detection in YouTube comments. *Soc Netw Anal Min*. 2022;12(1):75.
10. Chakravarthi BR. Multilingual hope speech detection in English and Dravidian languages. *Int J Data Sci Anal*. 2022;14(4):389-406.
11. Dar MA, Pushparaj J. Machine learning and deep learning approaches for accent recognition: a review. *IEEE Access*. 2025;13:51527-51550.
12. Dhanjal AS, Singh W. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools Appl*. 2024;83(8):23367-23412.
13. Fendji JLKE, Tala DC, Yenke BO, Atemkeng M. Automatic speech recognition using limited vocabulary: A survey. *Appl Artif Intell*. 2022;36(1):2095039.
14. Gandhi A, Ahir P, Adhvaryu K, Shah P, Lohiya R, Cambria E, *et al*. Hate speech detection: A comprehensive review of recent works. *Expert Syst*. 2024;41(8):e13562.
15. Hema C, Marquez FPG. Emotional speech recognition using cnn and deep learning techniques. *Appl Acoust*. 2023;211:109492.
16. Kumar Y, Koul A, Singh C. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools Appl*. 2023;82(10):15171-15197.
17. Luo Q. The improving effect of intelligent speech recognition System on english learning. *Adv Multimedia*. 2022;2022(1):2910859.
18. Ma P, Petridis S, Pantic M. Visual speech recognition for multiple languages in the wild. *Nat Mach Intell*. 2022;4(11):930-939.
19. Markl N. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*; 2022 Jun. p. 521-534.
20. Mehta H, Passi K. Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*. 2022;15(8):291.
21. Mukhamadiyev A, Khujayarov I, Djuraev O, Cho J. Automatic speech recognition method based on deep learning approaches for Uzbek language. *Sensors*. 2022;22(10):3683.
22. Mukhamadiyev A, Mukhiddinov M, Khujayarov I, Ochilov M, Cho J. Development of language models for continuous Uzbek speech recognition system. *Sensors*. 2023;23(3):1145.
23. Orken M, Dina O, Keylan A, Tolganay T, Mohamed O. A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci Rep*. 2022;12(1):8337.
24. Oruh J, Viriri S, Adegun A. Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*. 2022;10:30069-30079.
25. Seble H, Muluken S, Kafte T, Terefe F, Mekashaw G, Abiyot B, *et al*. Hate speech detection using machine learning: a survey. *AJSE*. 2023;20(1).
26. Sharrab YO, Attar H, Eljinini MAH, Al-Omary Y, Al-Momani WAE. Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions. *IEEE Access*. 2025;13:46925-46940.
27. Subramanian M, Sathiskumar VE, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and

sentiment analysis using machine learning and deep learning models. *Alexandria Eng J*. 2023;80:110-121.
28. Wang L. English speech recognition and pronunciation quality evaluation model based on neural network. *Sci Program*. 2022;2022(1):2249722.
29. Wang S. Recognition of English speech–using a deep learning algorithm. *J Intell Syst*. 2023;32(1):20220236.
30. William P, Gade R, Chaudhari RE, Pawar AB, Jawale MA. Machine learning based automatic hate speech recognition system. In: *2022 International conference on sustainable computing and data communication systems (ICSCDS)*; 2022 Apr. p. 315-318.
31. Xu H. Improving English speech recognition system accuracy using machine learning. In: *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*; 2024 Sep. p. 73-78.
32. Xu J, Li T. Application of multimodal NLP instruction combined with speech recognition in oral english practice. *Mobile Inf Syst*. 2022;2022(1):2262696.
33. Xu Y. English speech recognition and evaluation of pronunciation quality using deep learning. *Mobile Inf Syst*. 2022;2022(1):7186375.