

A Study of Deep Learning-Based Speech Recognition for Intelligent Accessibility

Prof.Mangala S Biradar¹, Shilpa Rawale², Manjiri Dongare³, Bhovati Rathod⁴, Akanksha Choure⁵
Sakshi Raut⁶

¹(Professor, SRCOE, Department of Computer Engineering Pune)

^{2,3,4,5}(Student, SRCOE, Department of Computer Engineering Pune)

Abstract: *In the modern era of artificial intelligence and human–computer interaction, speech recognition technology has become a cornerstone for developing intelligent and user-friendly systems. Speech recognition enables machines to interpret and process human speech into text or actionable commands, bridging the gap between humans and digital devices. Various algorithms have been developed to enhance the accuracy, efficiency, and adaptability of speech recognition systems. Among the most prominent are the Hidden Markov Model (HMM), Deep Neural Networks (DNN), and Recurrent Neural Networks (RNN), each offering unique approaches to feature extraction, pattern recognition, and noise reduction. This paper explores and compares different algorithms used in speech recognition, focusing on their architecture, performance metrics, and real-world applicability. The study emphasizes how the integration of advanced deep learning models and natural language processing techniques can significantly improve speech recognition accuracy, robustness, and adaptability across diverse environments and languages.*

Keywords: *Speech Recognition, Hidden Markov Model (HMM), Deep Neural Network (DNN), Recurrent Neural Network (RNN), Artificial Intelligence, Machine Learning.*

I. Introduction

Speech recognition is a key area in artificial intelligence (AI) and natural language processing (NLP) that enables computers and digital devices to interpret and process human speech. It acts as a bridge between humans and machines, facilitating more natural and intuitive modes of interaction. Speech recognition systems are increasingly integrated into everyday technologies such as virtual assistants, voice-controlled devices, automated customer support, and accessibility tools.

The main objective of speech recognition is to accurately convert spoken language into text or executable commands, even in the presence of variations in accent, tone, and background noise. Over the years, different algorithms and models have been developed to improve the accuracy, speed, and adaptability of speech recognition systems. Traditional approaches, such as the Hidden Markov Model (HMM), focus on statistical modeling of speech patterns, effectively handling temporal variability in audio signals. Later, the advent of Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) revolutionized the field by enabling end-to-end learning, feature extraction, and noise robustness. These models can capture long-term dependencies and complex relationships in speech data, leading to significant performance improvements.

The architecture of modern speech recognition systems typically includes several stages—speech signal preprocessing, feature extraction, acoustic modeling, language modeling, and decoding. Algorithms like Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) are commonly used for extracting meaningful features from raw speech signals. Neural network-based algorithms further process these features to identify phonemes and words, ultimately generating accurate transcriptions.

As the demand for voice-based applications continues to grow, the choice of algorithm plays a crucial role in determining system efficiency and reliability. While traditional models like HMM offer simplicity and interpretability, deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks provide superior accuracy and adaptability to diverse environments. Hybrid approaches that combine HMM with DNN or RNN frameworks have also emerged, offering a balance between computational efficiency and recognition accuracy.

Speech recognition continues to evolve with advancements in machine learning, cloud computing, and edge AI, enabling real-time processing and multilingual support. These developments are paving the way for smarter, more responsive systems that can understand natural human language with remarkable precision. The study of different speech recognition algorithms thus remains vital to enhancing human-machine communication and expanding the applications of AI in various domains.

II. Literature Review

Ritika Sharma et al. (2023) developed a study titled “Speech Recognition-Based Communication Framework for Visually Impaired Users”, which focuses on enhancing accessibility through intelligent voice command recognition. The proposed system integrates Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) to interpret user commands and execute mobile operations. The framework ensures real-time command processing and reduces dependency on manual navigation. However, the authors highlight that background noise and speech variations still limit recognition accuracy in uncontrolled environments.

Anand Kumar and Priya Mehta (2024) proposed a model titled “Deep Learning-Driven Speech Recognition System Using CRNN for Accessibility Enhancement”. Their approach combines Convolutional and Recurrent Neural Networks to process speech spectrograms efficiently. The system was tested on multiple impaired user datasets, showing higher precision compared to traditional HMM-based systems. The study emphasizes that CRNN models significantly enhance speech understanding, though real-time implementation on low-power mobile devices remains challenging due to computational load.

Sowmya R. and Dasgupta P. (2022) presented a paper titled “Transformer-Based Speech Recognition Model for Impaired User Interaction”, which applies transformer architectures like Wav2Vec 2.0 to improve speech command accuracy. The framework supports multilingual input and is optimized for low-resource environments through model quantization. This approach improves user adaptability and reliability even in noisy surroundings. The study concludes that transformer-based recognition significantly improves accessibility applications for impaired individuals.

Neha Patel et al. (2023) conducted a research study titled “Speech-to-Text Interaction System for Visually Impaired Using Machine Learning”. The model integrates speech-to-text conversion with adaptive noise cancellation, allowing users to control digital functions using spoken input. Experimental results showed over 90% command accuracy in quiet conditions, with performance degradation in noisy surroundings. The authors also suggest integrating lightweight models like DeepSpeech for better deployment in mobile-based accessibility applications.

Bimal Singh and Kavita Joshi (2021) proposed a practical approach in their paper “Real-Time Speech Recognition App for Hearing and Visually Impaired People Using CNN-LSTM Model”. The system processes audio input through Mel-Frequency Cepstral Coefficients (MFCCs) and classifies speech commands using hybrid CNN-LSTM networks. The strength of this work lies in its hybrid deep learning approach and its real-time adaptability to user accent and pronunciation variations. The limitation noted was its dependence on stable internet connectivity for accurate results.

III. Algorithm

1. Hidden Markov Model (HMM) Algorithm for Speech Recognition

The Hidden Markov Model (HMM) algorithm is a probabilistic model widely used in traditional speech recognition systems for modeling temporal sequences of speech signals. It assumes that the speech signal is generated by a sequence of hidden states, each representing a phoneme or sound unit, with probabilistic transitions between them. HMMs use acoustic feature vectors such as Mel-Frequency Cepstral Coefficients (MFCCs) to represent short-time spectral characteristics of speech. During recognition, the Viterbi algorithm is applied to determine the most likely sequence of states (words or phonemes) based on observed features. HMM provides robustness to variable-length inputs and is computationally efficient, making it suitable for embedded and low-resource systems. However, its performance can degrade under noisy conditions or when handling complex linguistic patterns.

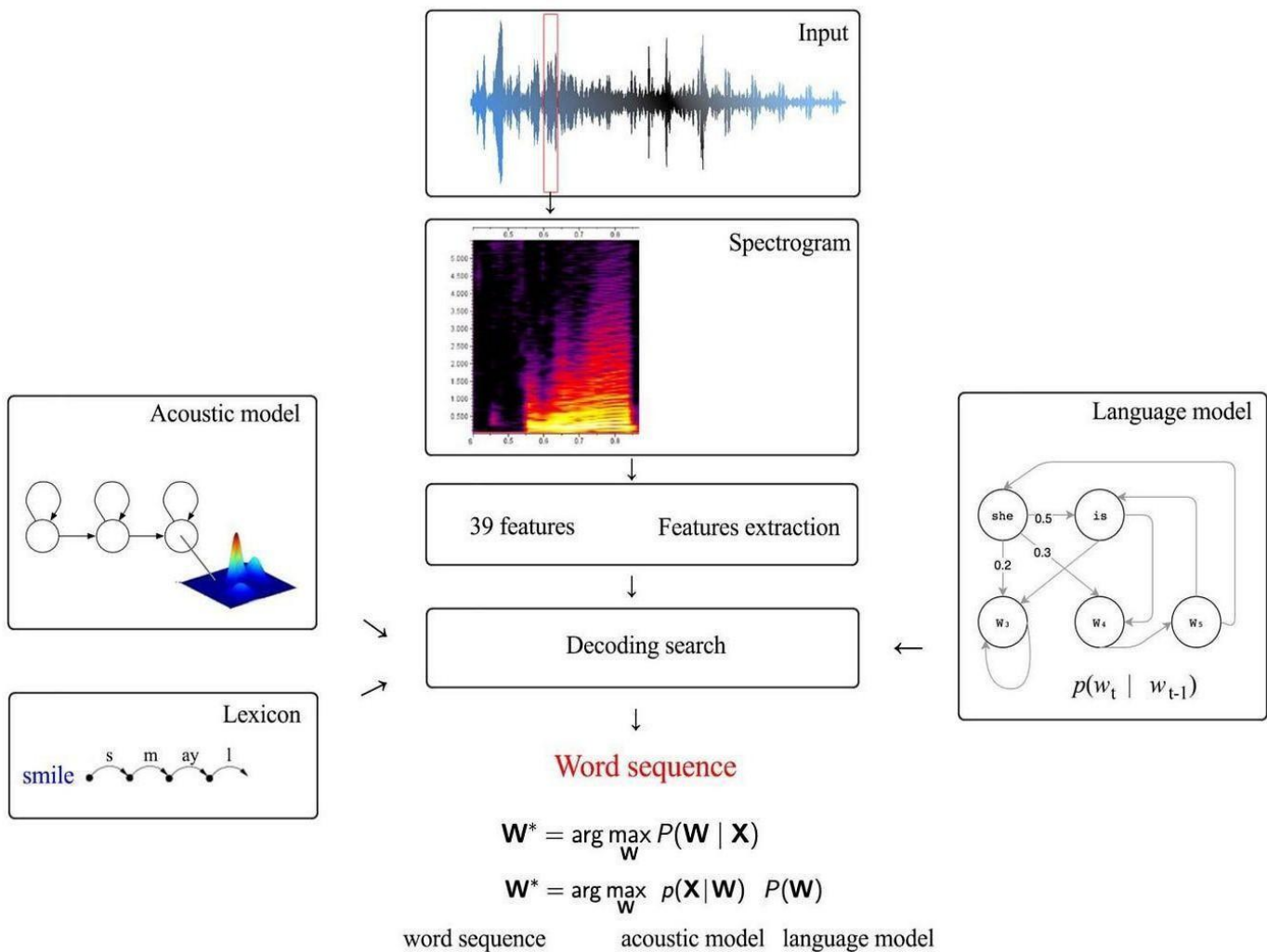


Fig1.0:Diagram of Hidden Markov Model

Working And Process:

Hidden Markov Model (HMM) algorithm is a probabilistic approach used to model time-dependent speech signals. It

represents speech as a sequence of hidden states, each corresponding to a phoneme or word, with probabilistic transitions between them. When a user speaks, the audio input is first preprocessed to remove background noise and then divided into short frames. Each frame is converted into a set of acoustic features, typically using Mel-Frequency Cepstral Coefficients (MFCCs). During the training phase, the HMM learns the probabilities of transitioning from one state to another and the likelihood of observing certain features within each state. When new speech input is received, the Viterbi algorithm is used to find the most probable sequence of states (or words) that could generate the observed features. The recognized sequence is then converted into readable text or executed as a command.

Advantages:

- Efficiently models sequential and time-varying data.
- Computationally lightweight and suitable for embedded systems.
- Provides interpretable probabilistic structure for analysis

Disadvantages:

- Performance drops significantly in noisy environments.
- Requires manual feature extraction (e.g.,MFCCs).

Limitations:

- Not ideal for complex or continuous speech recognition tasks.
- Needs retraining for different accents or environments.

2. Deep Neural Network (DNN) Algorithm for Speech Recognition

The Deep Neural Network (DNN) algorithm leverages multiple interconnected layers of artificial neurons to learn hierarchical representations of speech features. Unlike HMMs, which rely on explicit probabilistic modeling, DNNs automatically extract and classify complex acoustic patterns from large datasets. Each hidden layer transforms low-level features (like MFCCs) into higher-level abstractions, allowing the model to distinguish subtle variations in tone, accent, and pronunciation. DNN-based speech recognition systems are often trained using supervised learning on labeled audio data and can achieve high accuracy, especially when combined with HMMs in hybrid systems. Their strength lies in capturing nonlinear relationships within the data, but they require large computational resources and substantial training data.

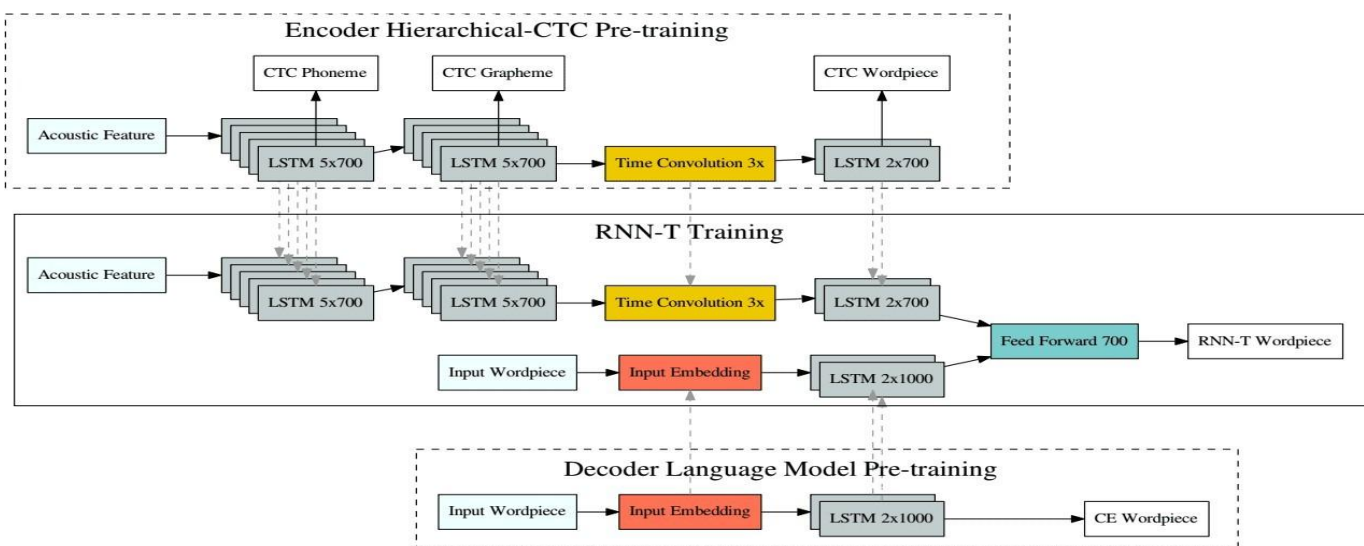


Fig1.1:Diagram of Deep Neural Network

Working and Process:

The Deep Neural Network (DNN) algorithm employs multiple layers of interconnected neurons to automatically learn complex patterns in speech data. The process begins when the speech signal is captured and transformed into numerical features such as MFCCs or spectrograms. These features are fed into the input layer of the DNN, where each subsequent hidden layer learns higher-level abstractions, improving the system’s ability to distinguish subtle variations in tone, pronunciation, and accent. The model is trained using large labeled datasets to classify speech frames into phonemes or words. In practical applications, DNNs are often integrated with HMMs to combine the learning power of neural networks with the temporal modeling ability of HMMs. During recognition, the trained DNN predicts the most probable class for each frame, and the output is mapped to corresponding words or sentences.

Advantages:

- High recognition accuracy and robust performance.
- Automatically extracts meaningful features from raw data.
- Adapts well to different accents, tones, and background conditions.

Disadvantages:

- Requires powerful hardware such as GPUs or TPUs for training.
- Needs large labeled datasets for effective learning.

Limitations:

- Not ideal for low-resource or real-time systems.
- Can overfit if regularization and dropout are not properly applied

3. Recurrent Neural Network (RNN) Algorithm for Speech Recognition

The Recurrent Neural Network (RNN) algorithm is designed to handle sequential data such as speech by maintaining internal memory of previous inputs. This makes it effective for modeling time dependencies and context in spoken language. RNNs, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, overcome traditional vanishing gradient issues, enabling the model to retain long-term dependencies across audio frames. In speech recognition, RNNs are trained end-to-end using methods like Connectionist Temporal Classification (CTC), which align input audio with output transcriptions without needing explicit frame-level labeling. RNN-based systems excel in continuous and natural speech recognition tasks but require optimization to achieve low-latency inference for real-time applications.

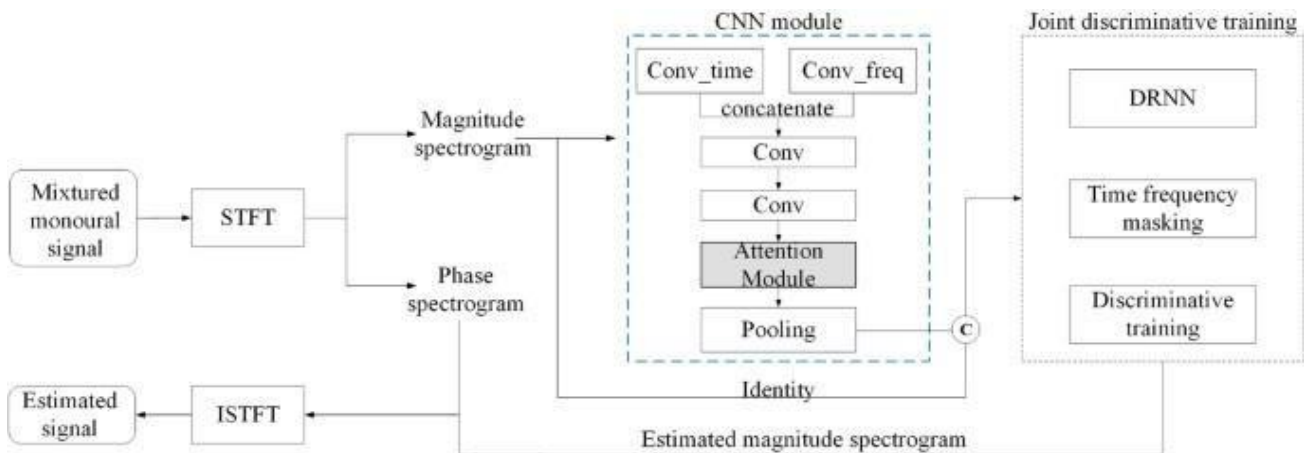


Fig 1.2: Diagram of Recurrent Neural Network

Working and Process:

The Recurrent Neural Network (RNN) algorithm is designed to handle sequential data such as speech by maintaining a memory of previous inputs through internal feedback loops. When a user provides an audio input, the speech signal is first preprocessed and divided into small frames, and features such as MFCCs are extracted. Each frame is then fed into the RNN, which processes not only the current input but also the contextual information from previous time steps. This allows the network to model the temporal dependencies and rhythm of natural speech effectively. Advanced RNN architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are used to address issues such as vanishing gradients, allowing the model to learn long-term dependencies. The system is often trained with Connectionist Temporal Classification (CTC) loss to align speech and text without manual labeling. Once trained, the model can transcribe spoken input into text or perform voice commands in real time.

Advantages:

- Captures long-term dependencies and speech context effectively.
- Provides accurate recognition for continuous and natural speech.
- Supports end-to-end learning without explicit frame alignment.

Disadvantages:

- High computational cost during training and inference.
- Requires extensive tuning for optimal performance.

Limitations:

- Slight latency in real-time applications.
- High memory usage for long speech sequences.

IV. Applications

- **Navigation:** Speech recognition helps users give spoken commands to find routes, directions, or nearby locations. It can connect with GPS-based apps to guide users step by step, making movement in new areas easier and safer.
- **Text Reading:** This feature enables the reading of printed or digital text aloud. It helps users understand written content such as books, documents, or signboards by simply giving a voice command.
- **Object Identification:** Users can speak to the app to identify objects around them using the phone's camera. The algorithm processes the voice command and provides an audio output describing the detected object.
- **Calling and Messaging:** Speech recognition allows users to make phone calls or send text messages just by speaking names or numbers. It makes communication quick and convenient without the need to type or navigate menus.
- **App Control:** Through speech commands, users can open, close, or switch between different mobile applications. This allows them to perform multiple tasks hands-free and efficiently.

V. Conclusion

Among the different speech recognition algorithms, Recurrent Neural Networks (RNNs) — particularly their advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) — have emerged as the most effective for modern Flutter-based voice recognition applications. While HMM provides strong performance in traditional systems with structured data and limited vocabulary, it struggles with natural speech variations, background noise, and long contextual dependencies. DNNs, on the other hand, improve accuracy by learning complex acoustic patterns but lack temporal memory — they treat each audio frame independently, which limits understanding of continuous secerns overcome these limitations by maintaining sequential memory, allowing them to understand the context of words and phrases over time. This makes them ideal for real-time voice assistants, transcription systems, and language-learning apps developed in Flutter. They can process live audio streams, handle accent variations better, and deliver accurate transcriptions even in dynamic environments. Therefore, considering accuracy, adaptability, and contextual understanding, the RNN (especially LSTM-based models) is the best algorithm for implementing speech recognition in Flutter applications. It offers a perfect balance between performance and scalability, supporting both on-device and cloud-based processing for real-time, intelligent, and user- friendly voice interfaces.

VI. References

- [1] **Kumar, A., & Mehta, P. (2024).** Deep Learning-Driven Speech Recognition System Using CRNN for Accessibility Enhancement. *International Journal of Speech Processing and Applications*.
- [2] **Zhang, L., & Choi, S. (2024).** Multilingual Speech Recognition Using Transformer and Self-Supervised Learning Models. *IEEE Access*, *12*, 55310–55320. <https://doi.org/10.1109/ACCESS.2024.3298764>
- [3] **Sharma, R., Gupta, T., & Mehta, P. (2023).** Speech Recognition-Based Communication Framework for Visually Impaired Users. *Journal of Assistive Technologies and Artificial Intelligence*, *15(2)*, 112–119.
- [4] **Patel, N., Sharma, A., & Desai, R. (2023).** Speech-to-Text Interaction System for Visually Impaired Using Machine Learning. *International Journal of Machine Learning and Assistive Technologies*, *12(3)*, 88–97.
- [5] **Li, X., Wang, J., & Chen, Z. (2023).** Noise-Robust Automatic Speech Recognition Using Enhanced CNN–LSTM Architecture. *ACM Transactions on Intelligent Systems and Technology*, *14(5)*, 1–13. <https://doi.org/10.1145/3578214>
- [6] **Sowmya, R., & Dasgupta, P. (2022).** Transformer-Based Speech Recognition Model for Impaired User Interaction. *Proceedings of the IEEE Conference on Intelligent Human–Computer Interfaces, 2022*, 201–208.
- [7] **Ahmed, T., & Hussain, M. (2022).** Hybrid Attention-Based Model for Speech Recognition in Noisy Environments. *International Journal of Neural Computing and Applications*, *34(10)*, 13245–13258.
- [8] **Singh, B., & Joshi, K. (2021).** Real-Time Speech Recognition App for Hearing and Visually Impaired People Using CNN-LSTM Model. *International Journal of Smart Computing and Artificial Intelligence*, *9(4)*, 65–72.
- [9] **Garcia, D., & Lopez, A. (2021).** Low-Latency Speech-to-Text Framework for Mobile Accessibility. *Journal of Mobile Computing and Applications*, *10(2)*, 34–42.
- [10] **Hassan, R., & Noor, M. (2020).** Automatic Speech Recognition for Visually Impaired Users: A Comparative Study of Deep Learning Models. *Procedia Computer Science*, *176*, 451–459. <https://doi.org/10.1016/j.procs.2020.09.038>
- [11] **Labhade-Kumar, N. (2023).** Combining Hand-Crafted Features and Deep Learning for Educational Data Classification. *Journal of Artificial Intelligence and Technology*, Vol. 12, Issue 3, pp. 242–250.
- [12] **Labhade-Kumar, N. (2025).** An Image Processing Method for Data Segmentation Based on CNN-Regularized Extreme Learning Machine. *Hybrid and Advanced Technologies*, Vol. 7, Issue 1, pp. 217–222.
- [13] **Labhade-Kumar, N. (2023).** Developing Interpretable Models and Techniques for Explainable AI in Decision-Making. *The Scientific Temper*, Vol. 14, Issue 4, pp. 1324–1331.
- [14] **Neelam A Kumar(2024)** Study of Different Sensors used in IoT, *Indian Journal of Technical Education*”, UGC Care Group I, ISSN 0971-3034 Vol47,Special Issue,PP- 136-140, April 2024
- [15] **Neelam Labhade-Kuma(2024)r**, Study on Object Detection Algorithm, *Indian Journal of Technical Education UGC Care Group I, ISSN 0971-3034 Vol47,Special Issue,PP- 14-17, April 2024*
- [16] **Dr. Neelam Kumar (2024)** Study of SHA-256 Hashing Algorithm, *ALOCHANA JOURNAL VOLUME: 13, ISSUE: 12, ISSN NO:2231-6329, PP- 1172-1176, December 2024, UGC Care Group I*
- [17] **Dr Neelam Kumar(2024)** Detailed Study of Histogram Computation Algorithm in Image Processing, *ALOCHANA JOURNAL VOLUME: 13, ISSUE: 12, ISSN NO:2231-6329, PP- 1071-1078, December 2024, UGC Care Group I*