As per UGC guidelines an electronic bar code is provided to secure your paper

International Journal for Modern Trends in Science and Technology
Volume 12, Special Issue 01, pages 399-405.
ISSN: 2455-3778 online
Available online at: http://ijmtst.com/ICRCSET2026.html
DOI: https://doi.org/10.5281/zenodo.19561888

Artificial Intelligence based Speech Recognition model

Konda Aswani Kumar Reddy, S. Sree Chandra

Department of Electronics and Communication Engineering, Chalapathi Institute of Technology, Abburi Ragavaiah Nagar, Mothadaka, Guntur, Andhra Pradesh, India

To Cite this Article
Konda Aswani Kumar Reddy & S. Sree Chandra (2026). Artificial Intelligence based Speech Recognition model. International Journal for Modern Trends in Science and Technology, 12(SI01), 399-405. https://doi.org/10.5281/zenodo.19561888

Article Info
Received: 02 March 2026; Revised: 01 April 2026; Accepted: 04 April 2026.

KEYWORDS

Artificial Intelligence (AI), Deep Neural Networks (DNNs), Speech Recognition, Human-Machine Interaction (HMI), Precision, Recall.

ABSTRACT

Speech Recognition is a kind of technology which allows the user to operate the electronic device through spoken word instead of using different tools such as keystrokes, button and keyboard etc. Advancements in Artificial Intelligence (AI) have significantly improved Human-Machine Interaction (HMI), especially with technologies that convert speech into executable actions. It is the process of enabling a computer to recognize and revert to the sounds produced in human speech. Speech recognition is also known as automatic speech recognition (ASR). This paper presents, Artificial Intelligence based Speech Recognition model. In this study, the dataset with 120h of audio which consist of sentences with a maximum of 15 words was used for the model training. Deep Neural Networks (DNNs) network is used in this paper for Speech Recognition. Precision, recall and WER (Word Error Rate) are the used parameters for performance analysis. The experimental results demonstrate that the proposed method successfully detects the speech signals and achieves seamless classification performance compared to other conventional speech recognition algorithms.

I. INTRODUCTION

Speech Recognition is a kind of technology which allows the user to operate the electronic device through spoken word instead of using different tools such as keystrokes, button and keyboard etc. Speech recognition software convert the words and phrases which is spoken by user into machine-readable format so that user can easily operate the device through speech [1].

Speech recognition which is also known as automatic speech recognition (ASR). The main objective of developing speech recognition is any people whether it is technical or non-technical can easily operate the device. As well as an illiterate which have no knowledge about device and its parts they can be operated it very easily [2]. Speech recognition is basically designed for a single user.

Today, human interaction with machines use devices like mouse and keyboards which depend much on hand movements, the speech technology can change the norms by allowing interaction via speech which is faster,

easy and comfort. Human speech perception starts with receiving speech waveform through the ears. The speech will enter membrane basilar situated in the inner middle ear at which the signal waveform will be analyzed and produced spectrum signal. The spectrum signal will enter neural transducer which converts the signal to neural activities at the ear's nerve [3]. The neural signal activities are translated into language code and the message will be sent to the brain for perception.

Research in speech processing and communication for the most part, was motivated by peoples desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years [4]. Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Avoinics, Automobile portal, speech transcription, Handicapped people (blind people) supermarket, railway reservations etc.

Advancements in artificial intelligence (AI) have significantly improved human-machine interaction (HMI), especially with technologies that convert speech into executable actions. automatic speech recognition (ASR) emerges as a leading communication technology in HMI, extensively utilized by corporations and service providers for facilitating interactions through AI platforms like chatbots and digital assistants [5]. Spoken language forms the core of these inter actions, emphasizing the necessity for sophisticated speech processing in AI systems tailored for ASR. This paper presents, Artificial Intelligence based Speech Recognition model. In this study, the dataset with 120h of audio was used for the model training. Finally, Precision, recall and WER (Word Error Rate) are the used parameters for performance analysis.

The remainder of this paper is organized as follows: section II presents literature survey, Section III explains the described Artificial Intelligence based Speech Recognition model. Results and discussions are presented in section IV, and finally, Section V concludes the study by summarizing the findings.

II. LITERATURE SURVEY

In [6] focuses on developing an automatic speech recognition system for English lectures, which involves summarizing the content and providing Japanese subtitles. Subtitling the entire audio of an English lecture could hinder comprehension and readability, so a summarization system is desired. By employing the DNN-HMM based speech recognition system, we achieved an 88% word accuracy for recognizing TED lecture speeches. Speech translation results showed a lower BLEU score of approximately 14% compared to text translation.

In [7] exploit a hierarchical Bayesian interpretation for language modeling, based on a nonparametric prior called Pitman–Yor process. This offers a principled approach to language model smoothing, embedding the power-law distribution for natural language. Experiments on the recognition of conversational speech in multiparty meetings demonstrate that by using hierarchical Bayesian language models, we are able to achieve significant reductions in perplexity and word error rate. In [8], two techniques of automatic speech recognition system training on noised speech are compared with technique of training on clean speech. The comparing has been made by means of speech recognition accuracy measure, with usage of fourteen kinds of noise. It is shown that training on noised speech allows reaching the 95% recognition accuracy for minimal signal-to-noise ratio 10 dB, whereas training on clean speech allows reaching the same recognition accuracy for minimal signal-to-noise ratio 20 dB.

In [9] focuses on the role of speech recognition technology in speech therapy for children with

pronunciation disorders. A simple web game was designed to improve pronunciation of phonemes through children's rhymes. To improve the recognition of children's speech, a model was trained on a children's speech corpus, as conventional systems are optimized for adults. The newly trained model significantly improved the accuracy of speech recognition. In [10] first constructs a comprehensive dataset for Vietnamese Audio-visual Speech Recognition (VASR). A ViAVSP-LLM speech recognition system consisting of an AV-HuBERT encoder and VinaLLaMA decoder is then proposed. Comparative experiments conducted using current audio-only speech recognition models show that the addition of visual data significantly improves the speech recognition accuracy.

III. AI BASED SPEECH RECOGNITION MODEL

The block diagram of Artificial Intelligence based Speech Recognition model is shown in Figure 1.

In this study, the dataset with 120h of audio was used for the model training. The dataset includes speech audio recordings which consist of sentences with a maximum of 15 words with a total length of approximately 120h. In addition, the dataset includes a large amount of different text for use in developing the language model. In addition, the dataset includes a large amount of different text for use in developing the language model. Over 90,650 utterances, 415,780 words, and 65,810 unique words that were included in the text corpus were collected, resulting in around 120 h of transcribed speech data.

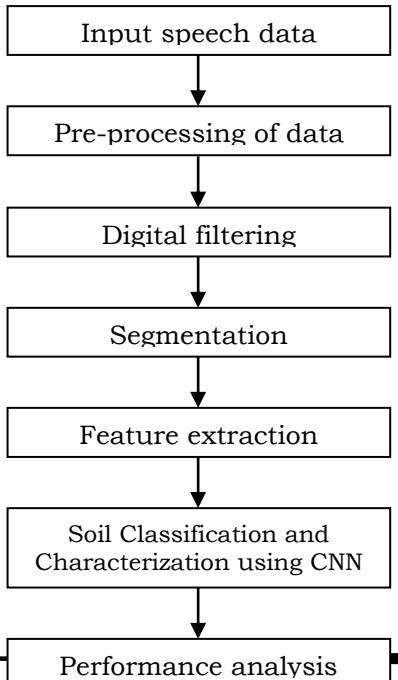

Figure 1. Block diagram of AI based Speech Recognition model

In addition to a typical, useful signal, various types of noise are present. As noise negatively affects the quality of speech recognition systems, dealing with noise is a press ing issue. Two types of digital filters are used to reduce the noise levels in the system: a line filter and an initial filter. A linear filter can be considered a combination of low- and high-frequency filters as it captures all low and high frequencies. Initial filtering is applied to minimize the impact of local disturbances on the characteristic markings that are used for subsequent identification. A speech signal must be passed through a low-pass filter for spectral alignment.

Segmentation is the process of dividing a speech signal into discrete, non-overlap ping fragments. The signal is usually divided into speech units such as sentences, words, syllables, phonemes, or even smaller phonetic units. The segmentation of recordings that contain the utterances of numerous speakers may consist of attributing pieces of utterances to particular speakers. The term “segment-stations” is sometimes used to refer to a division of the speech signal into frames prior to its parameterization.

Mel-frequency Cepstral coefficients (MFCC) is the most common method for extracting speech features. The human ear is a nonlinear system concerning how it perceives the audio signal. In order to cope with the change in frequency, the Mel-scale was developed to make a linear model of the human auditory system. Only frequencies in the range of [0,1] kHz can be transformed to the Mel-scale, while the remaining frequencies are considered to be logarithmic.

The features extracted by the model are then used to train the Deep Neural Networks (DNNs) to recognize speech. Deep Neural Networks (DNNs) have significantly advanced the field of automatic speech recognition (ASR) by enabling systems to learn complex patterns from large volumes of speech data. DNNs work by using a hierarchy of layers to model complex relationships between the input speech and the

corresponding text output. A DNN is a multi-layered neural network composed of an input layer, several hidden layers, and an output layer. In speech recognition, the input typically consists of acoustic features such as MFCC (Mel-Frequency Cepstral Coefficients), while the output represents phonemes, characters, or words.

Finally, Precision, recall and WER (Word Error Rate) are the used parameters for performance analysis. Deep Neural Networks (DNNs) have significantly detects the speech signals.

IV. Result Analysis

In this section, the performance of Artificial Intelligence based Speech Recognition model is evaluated. In this study, the dataset with 120h of audio which consist of sentences with a maximum of 15 words was used for the model training. Precision, recall and WER are the used parameters for performance analysis.

The precision is the ratio of the number of correctly predicted positive observations to the total number of predicted positive observations. Equation 1 represents the precision parameter.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

The recall is the ratio of the number of correctly predicted positive observations to the total number of observations in the actual class. It is expressed in equation 2,

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Where, TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives.

Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER). Word error rate is a common metric of the performance of a speech recognition. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as in equation 3:

$$WER = \frac{S + D + I}{N} \quad (3)$$

Where, S is the number of substitutions, D is the number of the deletions, I is the number of the insertions and N is the number of words in the reference.

The comparative performance analysis of described Artificial Intelligence based Speech Recognition model (DNN) with other models of Speech Recognition using Decision Tree (DT) and K-Nearest Neighbor (KNN) is represented in below Table 1.

Table 1: Comparative Performance Analysis

Speech Recognition model	Precision (%)	Recall (%)	WER (%)
DT	84	86	35
KNN	86	89	29
DNN	96	97	4

Figure 2 represents the Precision parameter comparative analysis for described Artificial Intelligence based Speech Recognition model (DNN) with other models of Speech Recognition using DT and KNN. X-axis represents the classification models and Y-axis represents percentage of parameter value. From figure 2 it is clear that precision of described model is high compared to other models. Comparative performance analysis of recall parameter for described Artificial Intelligence based Speech Recognition model (DNN) with other models of Speech Recognition using DT and KNN is represented in below Figure 3, in which Y-axis shows the percentage value and X-axis denotes classification models. Recall parameter of DNN model is high than other models.

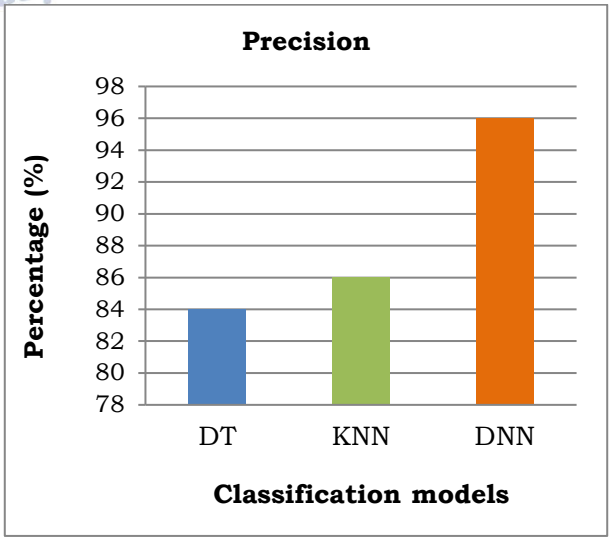

Figure 2. Comparative analysis for precision parameter

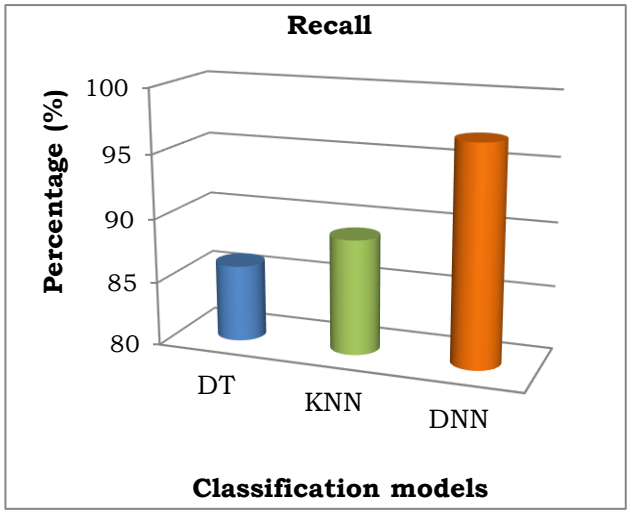

Figure 3. Comparative analysis for Recall parameter

WER parameter comparative analysis graphical representation is shown in Figure 4. X-axis represents the classification models and Y-axis represents percentage value. DNN model achieves less WER value than other two models.

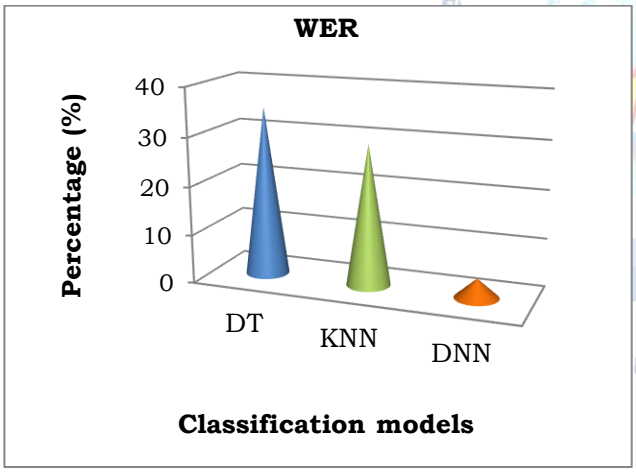

Figure 4. Comparative analysis for WER parameter

Therefore from overall results, described Artificial Intelligence based Speech Recognition model is efficient in terms of all parameters. Achieved parameters for described model are WET as 4%, precision as 96% and recall as 97%.

V. CONCLUSION

In this paper, Artificial Intelligence based Speech Recognition model is described. Speech Recognition System (SRS) is growing day by day and has unlimited applications. Speech recognition is an essential research aspect of speech signal processing and a vital human–computer interaction technique. Deep Neural Networks (DNNs) network is used in this paper for Speech Recognition. Deep Neural Networks (DNNs) have significantly advanced the field of automatic speech recognition (ASR) by enabling systems to learn complex patterns from large volumes of speech data. In this study, the dataset with 120h of audio which consist of sentences with a maximum of 15 words was used for the model training. Mel-frequency Cepstral coefficients (MFCC) is the most common method for extracting speech features. Precision, recall and WER are the used parameters for performance analysis. Achieved parameters for described model are WET as 4%, precision as 96% and recall as 97%. Therefore from overall results, described Artificial Intelligence based Speech Recognition model is efficient in terms of all parameters.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2313-2322, Dec. 2017, doi: 10.1109/TASLP.2017.2738559

[2] P. A. Asli and A. Zumbansen, "Performance of Speech Recognition Algorithms in Musical Speech used for Speech-Language Pathology Rehabilitation," 2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Jeju, Korea, Republic of, 2023, pp. 1-5, doi: 10.1109/MeMeA57477.2023.10171898.

[3] A. Buzo, H. Cucu, L. Petrică, D. Burileanu and C. Burileanu, "An Automatic Speech Recognition solution with speaker identification support," 2014 10th International Conference on Communications (COMM), Bucharest, Romania, 2014, pp. 1-4, doi: 10.1109/ICComm.2014.6866674.

[4] A. Touazi and M. Debyeche, "An SVD-based scheme for MFCC compression in distributed speech recognition system," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 2013, pp. 250-255, doi: 10.1109/ASRU.2013.6707738.

[5] J. S, A. H. G, M. S and M. A. Kumar, "Speech Enhancement Algorithm Analysis for a Reliable Speech Recognition System using Artificial Intelligence Methods," 2023 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICERCS57948.2023.10434226.

[6] K. Yamamoto, H. Banno, H. Sakurai, T. Adachi and S. Nakagawa, "A Study of Speech Recognition, Speech Translation, and Speech Summarization of TED English Lectures," 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 2023, pp. 451-452, doi: 10.1109/GCCE59613.2023.10315471.

[7] S. Huang and S. Renals, "Hierarchical Bayesian Language Models for Conversational Speech Recognition," in IEEE Transactions on

Audio, Speech, and Language Processing, vol. 18, no. 8, pp. 1941-1954, Nov. 2010, doi: 10.1109/TASL.2010.2040782

[8] A. Prodeus and K. Kukharicheva, "Training of automatic speech recognition system on noised speech," 2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kiev, Ukraine, 2016, pp. 221-223, doi: 10.1109/MSNMC.2016.7783147.

[9] S. Ondáš, J. Staš and R. Ševc, "Speech recognition as a supportive tool in the speech therapy game," 2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA), Zilina, Slovakia, 2024, pp. 1-4, doi: 10.1109/RADIOELEKTRONIKA61599.2024.10524060.

[10] T. -T. Duong, V. -M. Nguyen, H. -D. -K. Pham and T. -H. Le, "Vietnamese Automatic Speech Recognition Utilizing Audio and Visual Data," 2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Khanh Hoa, Vietnam, 2025, pp. 1-6, doi: 10.1109/MAPR67746.2025.11133884.

[11] S. Swarna *et al.*, "Optimized low-energy adaptive uneven clustering hierarchy for cognitive radio sensor networks," 2026.

[12] R. Thommandru *et al.*, "Millimetre wave self-isolated MIMO antenna with high isolation and radiation efficiency," in *Proc. IDCIoT*, IEEE, Jan. 2024, pp. 191–196.

[13] D. N. Ravikiran *et al.*, "Parametric facial landmark detection using active shape models."

[14] S. S. Vellela *et al.*, "Improving network security using intelligent ensemble techniques," in *Proc. AMATHE*, IEEE, May 2024, pp. 1–7.

[15] K. K. Kommineni and P. Ande, "Blockchain-driven key management and privacy-preserving data aggregation scheme for SDN-enabled MANETs," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 9, pp. 601–615, 2025.

[16] D. N. Ravikiran and C. V. Akhil, "A face recognition method for security applications in smart homes and cities."

[17] R. Saravanakumar *et al.*, "Analysis of circular waveguide antenna for 5G mid-band applications," in *Proc. ICACCS*, IEEE, Mar. 2024, pp. 560–566.

[18] B. Nancharaiah *et al.*, "Implementation and performance comparison of novel optimization approaches to counter starvation in wireless networks," *Int. J. Comput. Netw. Inf. Secur.*, vol. 17, no. 1, pp. 17–27, 2025.

[19] R. Thommandru, "Survey on MIMO antenna for 5G applications," 2022.

[20] S. Sree Chandra *et al.*, "Fruit classification based on shape, color, and texture using image processing techniques," *Int. J. Mod. Trends Sci. Technol.*, vol. 10, no. 3, pp. 100–107, 2024.

[21] R. Saravanakumar *et al.*, "Dual-band performance enhancement of square wheel antennas with FR4 substrate for sub-7 GHz applications," in *Proc. ACROSET*, IEEE, Sept. 2024, pp. 1–7.

[22] D. N. Ravikiran *et al.*, "Optimized advanced encryption standard (AES) with enhanced S-box and automated key generation."

[23] V. K. R. Devana *et al.*, "A novel compact MIMO-UWB antenna with improved isolation using parasitic elements," *Arab. J. Sci. Eng.*, 2025.

[24] S. Swarna and V. R. Kolluru, "Active channel selection by sensors using artificial neural networks," *Int. J. Eng. Educ. Res.*, vol. 12, no. 4, pp. 1466–1473, 2024.

[25] R. Thommandru, "Innovative meta ring array antenna design for Ka-band," 2004.

[26] C. H. Nagaraju *et al.*, "Assimilation of blockchain for augmenting IoT-based smart home security," in *Blockchain Technology for IoT and Wireless Communications*, CRC Press, 2023, pp. 79–87.

[27] R. Saravanakumar *et al.*, "An armor-mounted antenna with deflected ground for sub-6 GHz applications," in *Proc. ICRISST*, IEEE, Mar. 2024, pp. 1–7.

[28] D. N. Ravikiran and C. G. Dethe, "Improvements in routing algorithms to enhance lifetime of wireless sensor networks," *Int. J. Comput. Netw. Commun.*, vol. 10, no. 2, pp. 23–32, 2018.

[29] "Blockchain-enabled secure data aggregation for SDN-enabled ad-hoc networks," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 5, pp. 704–717, Jun. 2025.

[30] S. Swarna and V. R. Kolluru, "An intelligent data communication in IoT-based healthcare application using optimized routing protocol," *J. High Speed Netw.*, vol. 31, no. 2, pp. 159–179, 2025.

[31] R. Thommandru and R. Saravanakumar, "Performance analysis of circularly polarised MIMO antenna for wireless applications," in *Proc. ICICNIS*, IEEE, Dec. 2024, pp. 513–518.

[32] D. N. Ravikiran *et al.*, "Secure visual data processing: Image encryption and decryption through reversible logic gates in VLSI design," *Int. J. Mod. Trends Sci. Technol.*, vol. 10, no. 2, 2024.

[33] P. B. M. Krishna *et al.*, "Design of CMOS ring modulator by built-in thermal tuning," in *Cognitive Computing Models in Communication Systems*, 2022.

[34] R. Saravanakumar *et al.*, "Cross scoop fractal antenna design with notch at 15 degree for emerging applications at 5.2 GHz," in *Proc. RAEEUCCI*, IEEE, Apr. 2024, pp. 1–7.

[35] D. N. Ravikiran *et al.*, "IoT-based advanced automatic toll collection and vehicle detection system."

[36] B. Potti *et al.*, "Genetic algorithmic approach to mitigate starvation in wireless mesh networks," in *Proc. ICCT*, Springer, 2016.

[37] K. K. Kommineni and A. Prasad, "A review on privacy and security improvement mechanisms in MANETs," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 2, pp. 90–99, Dec. 2023.

[38] K. K. Kommineni and A. Prasad, "Enhancing data security and privacy in SDN-enabled MANETs through improved data aggregation protection and secrecy," *Wireless Pers. Commun.*, vol. 139, pp. 855–882, 2024.

[39] S. Sree Chandra *et al.*, "Verilog-based solution for multi-vehicle parking," *Int. J. Mod. Trends Sci. Technol.*, vol. 10, no. 2, pp. 394–400, 2024.

[40] D. N. Ravikiran *et al.*, "Reversible logic-based cryptography design for secure and efficient data processing."

[41] R. Thommandru, "Cost-effective circularly polarized MIMO antenna for Wi-Fi applications," Nov. 2024.

[42] Vellela, S. S., & Balamanigandan, R. (2022, December). Design of Hybrid Authentication Protocol for High Secure Applications in Cloud Environments. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 408-414). IEEE.

[43] Vellela, S. S., Balamanigandan, R., & Praveen, S. P. (2022). Strategic survey on security and privacy methods of cloud computing environment. *Journal of Next Generation Technology*, *2*(1).

[44] Vellela, S. S., & Balamanigandan, R. (2024). An efficient attack detection and prevention approach for secure WSN mobile cloud environment. *Soft Computing*, *28*(19), 11279-11293.

[45] Polasi, P. K., Vellela, S. S., Narayana, J. L., Simon, J., Kapileswar, N., Prabu, R. T., & Rashed, A. N. Z. (2026). Data rates transmission, operation performance speed and figure of merit

signature for various quadurature light sources under spectral and thermal effects. *Journal of Optics*, *55*(1), 633-643.

[46] Vellela, S. S., Rao, M. V., Mantena, S. V., Reddy, M. J., Vatambeti, R., & Rahman, S. Z. (2024). Evaluation of Tennis Teaching Effect Using Optimized DL Model with Cloud Computing System. *International Journal of Modern Education and Computer Science (IJMECS)*, *16*(2), 16-28.

[47] Praveen, S. P., Vellela, S. S., & Balamanigandan, R. (2024). SmartIris ML: harnessing machine learning for enhanced multi-biometric authentication. *Journal of Next Generation Technology (ISSN: 2583-021X)*, *4*(1).

[48] Vellela, S. S., Rao, M. V., Krishna, C. V. M., Rao, T. S., & Dasthavejula, R. (2026). Piezoelectric and Shape-Memory Materials for Actuators and Energy Harvesting in Mechanical, Electronics, and Biomedical Engineering Using AI-Based Design. In *Advanced Materials for Biomedical Devices* (pp. 195-206). CRC Press.

[49] Vellela, S. S., & Balamanigandan, R. (2024). Optimized clustering routing framework to maintain the optimal energy status in the wsn mobile cloud environment. *Multimedia Tools and Applications*, *83*(3), 7919-7938.

[50] Praveen, S. P., Nakka, R., Chokka, A., Thatha, V. N., Vellela, S. S., & Sirisha, U. (2023). A novel classification approach for grape leaf disease detection based on different attention deep learning techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(6), 2023.

[51] Vellela, S. S., & Balamanigandan, R. (2023). An intelligent sleep-awake energy management system for wireless sensor network. Peer-to-Peer Networking and Applications, 16(6), 2714-2731.