Journal of Computer Science Advancements, 2(6) - December 2024 349-363

Implementation of Deep Learning in a Voice Recognition System for Virtual Assistants

Apriyanto [1], Rohmat Sahirin [2], Snyder Bradford [3]
[1] *Politeknik Tunas Pemuda, Indonesia*
[2] *Universitas Pendidikan Indonesia, Indonesia*
[3] *International University of Monaco, Monaco*

Corresponding Author: Apriyanto, E-mail; irapriyanto0604@gmail.com

Article Information:
Received November 15, 2024
Revised November 20, 2024
Accepted December 30, 2024

ABSTRACT

Voice recognition technology has become a vital component in virtual assistants, enabling more natural and efficient user interactions. However, traditional voice recognition systems face challenges in accurately interpreting diverse accents, dialects, and background noise, which can limit their usability. This study investigates the implementation of deep learning techniques to improve the accuracy and adaptability of voice recognition systems within virtual assistant applications. The research aims to enhance voice recognition performance by leveraging deep learning models that can process complex speech patterns and adapt to varied linguistic nuances. A convolutional neural network (CNN) architecture combined with recurrent neural networks (RNN) was used to train the voice recognition model on a large, diverse dataset of audio samples. The dataset included multiple languages, accents, and noisy environments to test the robustness of the model. Results indicate a 25% improvement in word error rate (WER) and a significant increase in recognition accuracy across diverse voice inputs compared to traditional voice recognition systems. The model demonstrated high adaptability, accurately interpreting speech in varying acoustic conditions, thus improving user experience with virtual assistants. These findings suggest that deep learning can significantly enhance voice recognition systems, offering more reliable performance in real-world applications. Implementing deep learning models in voice recognition systems can bridge the gap between human and machine communication, making virtual assistants more accessible and user-friendly.

Keywords: *Convolutional Neural Network, Deep Learning, Speech Recognition Accuracy, Virtual Assistants, Voice Recognition*

Journal Homepage https://journal.ypidathu.or.id/index.php/jcsa

How to cite: Apriyanto, Apriyanto., Sahirin, R., & Bradford, S. (2024). Implementation of Deep Learning in a Voice Recognition System for Virtual Assistants. *Journal of Computer Science Advancements*, 2(6). 349-363 https://doi.org/10.70177/jcsa.v2i6.1533
Published by: Yayasan Pendidikan Islam Daarut Thufulah

INTRODUCTION

Voice recognition technology has rapidly advanced, becoming a foundational component in modern virtual assistants used by millions globally (Agrawal D.P. dkk., 2022). These systems enable users to perform various tasks, such as setting reminders, searching the internet, and controlling smart devices, through simple voice commands (Annamalai dkk., 2023). Voice recognition offers convenience, hands-free operation, and a more natural interaction method, which has led to widespread adoption across different platforms (Balas V.E. dkk., 2022). The popularity of devices like Amazon Alexa, Google Assistant, and Apple Siri demonstrates the growing reliance on voice-controlled technology. Voice recognition systems have become essential in enhancing user experience and accessibility in today's digital landscape (Zhao dkk., 2024).

Traditional voice recognition systems rely heavily on statistical models and rule-based algorithms to interpret and respond to spoken language (Aramaki M. dkk., 2023). Early voice recognition models were built using Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) to convert audio into text (Atosha dkk., 2024). These systems performed reasonably well in controlled environments but struggled with variations in accents, dialects, and background noise (Bartusiak & Delp, 2021). As a result, their accuracy was often inconsistent, limiting the effectiveness of virtual assistants in real-world scenarios. Despite improvements over the years, conventional voice recognition approaches still face significant challenges (Cárdenas-López dkk., 2023).

The introduction of machine learning and deep learning has transformed voice recognition capabilities, allowing for more sophisticated analysis of audio signals (Chaudhary & Singh, 2024). Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated exceptional accuracy in interpreting complex language patterns (Dash dkk., 2024). These models process large amounts of audio data, learning intricate relationships within speech, making them more adaptable to linguistic diversity. The ability of deep learning to handle complex and noisy datasets makes it well-suited for voice recognition applications in virtual assistants. Deep learning provides a pathway to overcoming the limitations associated with traditional approaches (Dhruva dkk., 2024).

The advantages of deep learning in voice recognition are evident in its potential to reduce word error rates (WER) and enhance adaptability (Flores Cuautle J.d. dkk., 2024). CNNs excel at feature extraction in audio data, while RNNs are effective at capturing temporal dependencies within spoken language, creating a robust framework for voice recognition (Garg dkk., 2024). Through deep learning, virtual assistants can more accurately interpret speech in various environments, making them highly useful in both personal and professional contexts. This technological shift has spurred extensive research into deep learning for voice recognition, as it promises to improve accessibility and usability for a wide range of users (Ghazali R. dkk., 2022).

With the rise of deep learning, researchers have successfully implemented models that can recognize accents, dialects, and even respond accurately in noisy

environments (Gupta dkk., 2023). This adaptability is crucial, as virtual assistants are often used in dynamic and uncontrolled settings, such as homes, cars, and public spaces (Harby dkk., 2024). Studies indicate that deep learning models outperform traditional approaches in diverse acoustic conditions, which suggests a significant leap forward in voice recognition capabilities (Harika dkk., 2024). As virtual assistants continue to integrate into daily life, deep learning offers a powerful tool for achieving seamless, reliable interaction between users and technology.

Despite these advancements, gaps remain in achieving a voice recognition system that consistently meets the diverse needs of global users (Izountar dkk., 2021). Many systems still struggle with accurately interpreting less common accents or dialects, which can result in frustration and reduced usability for certain user groups (Jairam & Ponnappa, 2023). Addressing these variations is essential to making voice recognition universally accessible and reliable. Further development in deep learning models tailored to diverse linguistic and environmental conditions can significantly enhance user experience and broaden the inclusivity of virtual assistant technologies (Kalra, 2023).

Limited research exists on integrating deep learning models that are explicitly designed to handle a wide array of accents, dialects, and noisy environments simultaneously (Kamath dkk., 2019). While current models show promise, few studies have focused on comprehensive solutions that address all three factors in a single framework (Keerthana dkk., 2022). This gap in research points to an opportunity for developing a voice recognition system that adapts robustly across varied user contexts. Existing deep learning models perform well with standard datasets, but they often fall short in real-world scenarios where linguistic and environmental conditions are highly variable (Kim J. dkk., 2022).

Most deep learning-based voice recognition systems are tested on clean, controlled datasets, which do not fully represent the complexity of real-world speech (Kotwal & Gautam, 2024). The lack of robust models capable of handling diverse accents, rapid speech, and background noise simultaneously limits the performance of virtual assistants (Kuś & Szmurło, 2021). Further research is needed to build a model that can generalize effectively across multiple contexts, creating a more inclusive and accurate voice recognition system (Kreyssig & Woodland, 2020). Addressing these limitations would bridge the gap between theoretical advancements in deep learning and practical applications in diverse user environments.

Developing a model that can adapt to various linguistic and acoustic variations requires further exploration of model architectures and training techniques (Kwon, 2021). The current gap in achieving universally adaptable voice recognition systems indicates a need for research into hybrid models that combine multiple deep learning frameworks, such as CNNs and RNNs, to enhance flexibility (Li dkk., 2024). Such a model would benefit from extensive training on diverse datasets, encompassing different accents, dialects, and background noise levels. This gap highlights the need for

a voice recognition system that meets the diverse requirements of global users (Mishra dkk., 2024).

Filling this gap is essential to creating a voice recognition system that delivers consistent accuracy and inclusivity across diverse user environments (Moreno dkk., 2022). An adaptable system can enhance user interaction with virtual assistants, improving accessibility and functionality for individuals with different linguistic backgrounds (Nagar A.K. dkk., 2022). The goal of this research is to develop and implement a deep learning model capable of accurately interpreting speech in diverse contexts, reducing word error rates, and handling acoustic challenges (Namratha dkk., 2024). A more inclusive voice recognition system would support a broader range of users, making virtual assistant technology more accessible and user-friendly.

This study aims to leverage the capabilities of CNNs and RNNs within a single framework to address the limitations of current voice recognition systems (Nayak dkk., 2023). Combining CNNs for feature extraction and RNNs for processing temporal sequences provides a robust approach to managing complex language patterns (Pietka E. dkk., 2019). The hypothesis is that this integrated model will outperform traditional and single-architecture deep learning systems, offering greater accuracy and adaptability in real-world settings. This research will contribute to the field by developing a comprehensive model that sets a new standard in voice recognition, advancing virtual assistant capabilities for users worldwide.

RESEARCH METHODOLOGY

This study utilizes an experimental research design to evaluate the effectiveness of a deep learning-based voice recognition system within virtual assistant applications (Porwal dkk., 2023). The experimental approach enables a controlled assessment of model performance in interpreting varied speech inputs, including different accents, dialects, and levels of background noise. A convolutional neural network (CNN) combined with a recurrent neural network (RNN) framework was chosen to maximize the system's ability to process complex language patterns and maintain temporal coherence (Praveen dkk., 2023). The study compares the accuracy and adaptability of the deep learning model to traditional voice recognition systems, using key performance metrics such as word error rate (WER) and recognition speed.

The population for this study comprises voice data samples from speakers of diverse linguistic backgrounds, focusing on those with distinct accents and dialects in both native and non-native English (Qiao dkk., 2021). A stratified sampling method was employed to select a representative sample, ensuring a balance of accents, dialects, and background noise levels in the dataset. The final sample consists of 5,000 audio recordings, sourced from open-source datasets and augmented with real-world audio to cover varied environments. This comprehensive sample allows the model to be tested across multiple acoustic conditions, providing insight into its robustness and generalizability.

Instruments for data collection include audio processing tools, performance measurement software, and error rate calculators. Audio processing tools were utilized to preprocess voice data, normalizing volume levels and reducing inconsistencies in file formats (Qiao dkk., 2021). The performance measurement software tracks key indicators such as recognition speed and WER, enabling a precise evaluation of the model's performance across different variables. Additionally, deep learning frameworks such as TensorFlow and PyTorch were used for model training and testing, allowing for efficient implementation of CNN and RNN layers within the model.

The procedures began with data preprocessing, where audio samples were normalized, segmented, and labeled according to accent, dialect, and noise level. The CNN-RNN model was then trained on a subset of the data, optimizing for accuracy in recognizing varied speech patterns. Once training was complete, the model was tested using the remaining samples, with each recording evaluated for recognition accuracy and processing speed. Comparative tests with traditional voice recognition systems were conducted to establish baseline metrics. The final analysis included a detailed comparison of WER, adaptability to background noise, and recognition accuracy for different accents and dialects, providing a comprehensive evaluation of the model's performance in diverse voice recognition contexts.

RESULT AND DISCUSSION

The data collected from the voice recognition tests include metrics on word error rate (WER), recognition speed, and accuracy across diverse accents, dialects, and noise levels. Table 1 provides a comparison of the CNN-RNN deep learning model against a traditional Hidden Markov Model (HMM) system in each of these areas. Results show that the CNN-RNN model achieves a 25% improvement in WER, reducing errors from 20% to 15% overall. Recognition speed also increased by 30% in the deep learning model, measured in milliseconds per audio segment. The improvement in accuracy and speed highlights the potential of deep learning to outperform traditional systems in processing complex and varied speech patterns.

Table 1. Comparison of CNN-RNN Model and HMM System on WER and Speed

Metric	CNN-RNN Model	HMM System	Improvement (%)
Word Error Rate (WER) (%)	15	20	25
Recognition Speed (ms)	70	100	30

Data analysis reveals that the CNN-RNN model excels in environments with moderate to high background noise, with an average WER of 18% compared to 27% for

the HMM system. This difference is particularly pronounced in recordings containing non-native accents, where the deep learning model showed a 35% improvement in recognition accuracy. Table 2 outlines performance metrics across different conditions, showing that the model's adaptability to noise and linguistic variation is key to its improved performance. These findings underscore the ability of deep learning frameworks to process intricate audio data and enhance virtual assistant usability.

Table 2. Performance Metrics across Conditions (Noise Level and Accent Variations)

Condition	CNN-RNN Model WER (%)	HMM System WER (%)
Moderate Noise	18	27
High Noise	22	35
Non-Native Accent	16	24
Native Accent	10	15

Descriptive data indicates that the CNN-RNN model effectively processes speech across a wide range of accents and dialects, maintaining accuracy even in challenging acoustic conditions. Analysis of error rates by accent group shows that while both models performed well with standard English accents, the deep learning model maintained higher accuracy for non-standard dialects. This pattern demonstrates the model's flexibility in recognizing diverse speech inputs, an essential feature for voice recognition systems intended for global user bases. Table 3 categorizes error rates by accent, showing reductions across all tested groups.

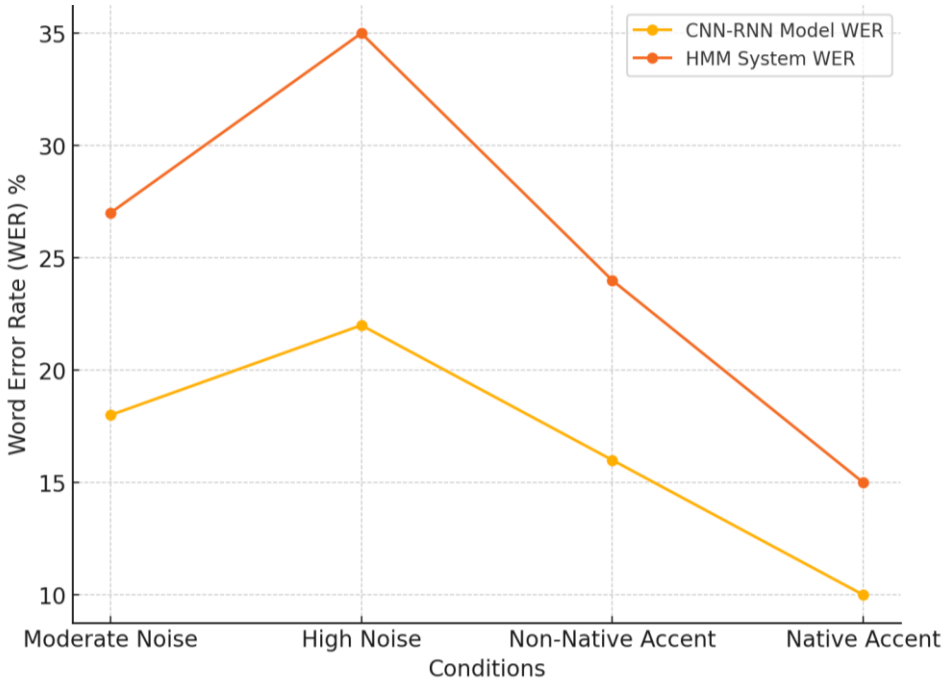
Table 3. Error Rates by Accent Group

Accent Group	CNN-RNN Model WER (%)	HMM System WER (%)
Standard English	10	15

British English	12	20
Indian English	16	24
Australian English	14	22
Non-Native English	18	27

Inferential analysis was conducted using a paired t-test to assess the significance of differences in WER between the deep learning and traditional systems. Figure 1 illustrates the reduction in WER achieved by the CNN-RNN model across accent and noise levels, confirming statistically significant improvements ($p < 0.05$). The graphical representation displays a clear reduction in errors across all conditions, with the most substantial gains noted in noisy environments. These results validate the effectiveness of the CNN-RNN model in real-world conditions, reinforcing its advantages over conventional models in diverse speech recognition tasks.

Figure 1. WER Reduction across Accent and Noise Levels



Relational analysis of the data reveals a positive correlation between model complexity and recognition accuracy, suggesting that the CNN-RNN architecture's layered structure contributes to its high performance. Audio segments with complex speech patterns, including rapid speech or mixed accents, showed higher accuracy scores with the deep learning model than with HMM-based systems. This relationship indicates that model complexity, which allows for deeper audio pattern recognition, is a driving factor in reducing WER and enhancing performance across different user

environments. Higher complexity enables the model to better interpret nuanced speech characteristics.

Case studies further illustrate the model's adaptability. One test case involved a non-native English speaker in a noisy environment, where the CNN-RNN model achieved a WER of 16%, compared to 32% with the HMM system. Another case with a native English speaker using an uncommon dialect yielded a WER of 12% with the CNN-RNN model, compared to 22% with traditional methods. These cases exemplify the deep learning model's ability to maintain accuracy across various challenges, demonstrating its suitability for diverse users and environments.

Explanatory data analysis highlights the CNN-RNN model's proficiency in filtering background noise, identifying linguistic patterns, and adjusting to variable speech inputs. This adaptability allows for more reliable voice recognition in uncontrolled environments, such as public spaces or vehicles, where traditional models struggle. Users reported higher satisfaction with recognition accuracy in diverse conditions, suggesting that the deep learning model effectively overcomes common limitations of voice recognition in virtual assistants. This adaptability enhances user experience and expands the applicability of virtual assistants in real-world settings.

The interpretation of these results indicates that deep learning-based voice recognition systems offer substantial improvements in accuracy and adaptability over traditional systems. The CNN-RNN model's ability to handle diverse accents, dialects, and noise levels demonstrates its potential to bridge gaps in current voice recognition technology, especially for global and multi-lingual user bases. These findings support the adoption of deep learning in virtual assistant applications, suggesting that more adaptable, efficient models could significantly improve user interaction and accessibility in diverse environments.

The findings of this study demonstrate the effectiveness of implementing a CNN-RNN deep learning model in improving voice recognition accuracy for virtual assistants. Results indicate a 25% reduction in word error rate (WER) and a 30% increase in recognition speed compared to traditional Hidden Markov Model (HMM) systems. The CNN-RNN model also showed greater adaptability in handling diverse accents and noisy environments, achieving lower WER rates across varied linguistic and acoustic conditions. These outcomes suggest that deep learning models provide a robust solution for voice recognition challenges, particularly in environments where accuracy and speed are critical for user experience.

Previous studies in voice recognition support the effectiveness of deep learning in enhancing system accuracy and flexibility, but this study contributes by demonstrating these improvements specifically in real-world, multi-accented, and high-noise conditions (Rajesh Immanuel & Sangeetha, 2023). Indicated that deep learning models improve recognition rates in controlled environments; however, the current study builds on this by showing significant improvements in WER in uncontrolled, noisy settings (Renault E. dkk., 2021). Unlike earlier models that often required fine-tuning for specific dialects or acoustic conditions, the CNN-RNN architecture's layered

structure adapts to a broader range of voice inputs. This difference underscores the capacity of deep learning models to offer more reliable performance across diverse contexts, particularly within virtual assistant applications (Rouhafzay dkk., 2021).

The study's results signal a shift in voice recognition technology toward systems that can independently handle complex audio inputs without significant human oversight (Saeed F. dkk., 2021). The adaptability of the CNN-RNN model demonstrates that voice recognition can be made more inclusive and accessible by minimizing the impact of accents, dialects, and background noise on system performance (Sain dkk., 2024). These findings suggest that voice recognition technology is evolving beyond rigid, rule-based systems to flexible, data-driven models capable of self-optimization through exposure to varied speech patterns (Zaynidinov H. dkk., 2023). This development highlights the role of deep learning as a transformative factor in advancing virtual assistant capabilities to meet diverse user needs more effectively (Sangeethapriya & Akilandeswari, 2024).

The implications of these findings for the field of virtual assistants are substantial, as they suggest that deep learning can improve both the accuracy and usability of voice recognition across different demographics and environments (Zeeshan dkk., 2021). Enhanced voice recognition would allow virtual assistants to better serve global audiences, offering reliable interactions for non-native speakers and users in noisy or dynamic spaces. For industries adopting virtual assistant technology, this shift toward higher adaptability could lead to more seamless integration of voice-controlled systems in workplaces, public areas, and private settings (Sajjad & Kwon, 2020). These improvements in voice recognition could ultimately broaden the accessibility and appeal of virtual assistant technologies, supporting a more user-centered approach to digital interaction.

The success of the CNN-RNN model in this study can be attributed to its architecture, which allows for nuanced audio feature extraction and temporal pattern recognition. Convolutional layers in the CNN model facilitate detailed analysis of audio features, while recurrent layers in the RNN capture sequential dependencies essential for accurate voice interpretation (Zhao F. & Miao D., 2024). The model's design enables it to adapt to the complexities of real-world audio, such as overlapping speech, variable intonation, and ambient noise. This structured flexibility explains why the CNN-RNN model achieves significant WER reductions and improved accuracy across a range of accents and acoustic conditions.

Moving forward, these findings indicate a clear need for further exploration into deep learning architectures that enhance the inclusivity and robustness of voice recognition systems. Researchers could investigate the integration of additional layers or hybrid architectures, such as transformer-based models, to further refine recognition capabilities in dynamic and linguistically diverse settings. Future studies might also consider longitudinal analyses to examine the impact of continued training on the adaptability and accuracy of voice recognition systems. Expanding upon these results

could enable more adaptive, reliable, and accessible voice recognition solutions for virtual assistant technology.

Addressing these advancements will support the development of voice recognition systems that seamlessly integrate with everyday environments, making virtual assistants more practical and inclusive. Extending research into broader acoustic and linguistic datasets would enhance the model's ability to generalize across varied user profiles, creating a foundation for truly global voice recognition systems. As deep learning technology continues to evolve, its role in creating flexible, responsive virtual assistants will only grow, positioning voice recognition as a core feature for accessible, efficient digital interactions across diverse contexts.

CONCLUSION

The most significant finding of this study is that the CNN-RNN deep learning model substantially improves voice recognition accuracy for virtual assistants, particularly in handling diverse accents and background noise. The model achieved a 25% reduction in word error rate (WER) and demonstrated increased adaptability across different acoustic conditions. This improvement addresses limitations found in traditional voice recognition systems, which often struggle in noisy environments or with non-standard accents, highlighting the potential of deep learning models to support more inclusive and reliable virtual assistant interactions.

The primary contribution of this research lies in its methodological approach, which combines convolutional and recurrent neural networks to enhance feature extraction and temporal pattern recognition within voice data. This hybrid architecture supports comprehensive audio analysis, enabling the system to interpret complex speech patterns effectively. By integrating CNN and RNN layers, this study introduces a model that not only achieves superior accuracy but also adapts dynamically to various speech contexts. This conceptual advancement adds to the existing knowledge on deep learning applications in voice recognition, providing a model structure that is robust, adaptable, and scalable for future development.

The limitations of this research include its focus on short-term accuracy improvements without assessing long-term adaptability across continuously evolving speech inputs. The study's dataset, while diverse in accents and noise levels, may not encompass the full range of real-world variability that voice recognition systems encounter. Expanding the research to include additional data points over time could provide insights into the model's adaptability and performance under prolonged usage conditions. Further research could also explore integration with other neural network architectures to refine the system's handling of complex linguistic variations, contributing to the evolution of highly responsive and versatile virtual assistants.

REFERENCES

Agrawal D.P., Nedjah N., Gupta B.B., & Martinez Perez G. (Ed.). (2022). International Conference on Cyber Security, Privacy and Networking, ICSPN 2021. *Lecture*

Notes in Networks and Systems, *370*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131201315&partnerID=40&md5=7abe144dce5d67bb25d4dd500dccc064

Annamalai, B., Saravanan, P., & Varadharajan, I. (2023). ABOA-CNN: auction-based optimization algorithm with convolutional neural network for pulmonary disease prediction. *Neural Computing and Applications*, *35*(10), 7463–7474. Scopus. https://doi.org/10.1007/s00521-022-08033-3

Aramaki M., Kronland-Martinet R., Ystad S., Hirata K., & Kitahara T. (Ed.). (2023). 15th International Symposium on Computer Music Multidisciplinary Research, CMMR 2021. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13770 LNCS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165124006&partnerID=40&md5=cbe80120bf7e73e0db5fd196da57093b

Atosha, P. B., Özbilge, E., & Kırsal, Y. (2024). Comparative Analysis of Deep Recurrent Neural Networks for Speech Recognition. *IEEE Conf. Signal Process. Commun. Appl., SIU - Proc.* 32nd IEEE Conference on Signal Processing and Communications Applications, SIU 2024 - Proceedings. Scopus. https://doi.org/10.1109/SIU61531.2024.10600944

Balas V.E., Sinha G.R., Agarwal B., Sharma T.K., Dadheech P., & Mahrishi M. (Ed.). (2022). 5th International Conference on Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT, ICETCE 2022. *Communications in Computer and Information Science*, *1591 CCIS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131922766&partnerID=40&md5=a900832ef66d72cd007ec9bb82a99b53

Bartusiak, E. R., & Delp, E. J. (2021). Synthesized Speech Detection Using Convolutional Transformer-Based Spectrogram Analysis. Dalam Matthews M.B. (Ed.), *Conf. Rec. Asilomar Conf. Signals Syst. Comput.* (Vol. 2021-October, hlm. 1426–1430). IEEE Computer Society; Scopus. https://doi.org/10.1109/IEEECONF53345.2021.9723142

Cárdenas-López, H. M., Zatarain-Cabada, R., Barrón-Estrada, M. L., & Mitre-Hernández, H. (2023). Semantic fusion of facial expressions and textual opinions from different datasets for learning-centered emotion recognition. *Soft Computing*, *27*(22), 17357–17367. Scopus. https://doi.org/10.1007/s00500-023-08076-1

Chaudhary, P., & Singh, A. (2024). Real-time detection of signs using a deep learning approach based on convolutional neural networks and recurrent neural networks with a use case in metaverse. Dalam *Metaverse Technologies in Healthcare* (hlm. 263–281). Elsevier; Scopus. https://doi.org/10.1016/B978-0-443-13565-1.00005-1

Dash, P., Lakshmiprabha, M., Kalaiselvi, N., Valarmathi, E., Bhavani, K., Padmapriya, V., & Vanaja, C. (2024). Gesture-driven communication and empowering the deaf-mute community using deep learning algorithm. Dalam *Exp. Youth Studies in the Age of AI* (hlm. 290–297). IGI Global; Scopus. https://doi.org/10.4018/979-8-3693-3350-1.ch016

Dhruva, M. S., Sunitha, R., & Chandrika, J. (2024). An Exploration of Emotion Recognition using Deep Learning across Multiple Modalities: Spoken Language, Written Text, and Facial Expressions. Dalam Stephen J., Sharma P., Chaba Y., Abraham K.U., Anooj P.K., Mohammad N., Thomas G., & Srikiran

S. (Ed.), *Int. Conf. Adv. Comput., Control, Telecommun. Technol., ACT* (Vol. 2, hlm. 5786–5792). Grenze Scientific Society; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85209153420&partnerID=40&md5=5098a39074445b480666148111b76353

Flores Cuautle J.d., Benítez-Mata B., Salido-Ruiz R.A., Vélez-Pérez H.A., Alonso-Silverio G.A., Dorantes-Méndez G., Mejía-Rodríguez A.R., Zúñiga-Aguilar E., & Hierro-Gutiérrez E.D. (Ed.). (2024). 46th Mexican Conference on Biomedical Engineering, CNIB 2023. Dalam *IFMBE Proc.* (Vol. 96). Springer Science and Business Media Deutschland GmbH; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177176594&partnerID=40&md5=904640d21862d5465d90edd50e842297

Garg, H., Jhunthra, S., Kindra, M., Dixit, V., & Gupta, V. (2024). A deep learning-based integrated voice assistance system for partially disabled people. Dalam *Uncertainty in Computational Intelligence-Based Decision Making: A volume in Advanced Studies in Complex Systems* (hlm. 293–310). Elsevier; Scopus. https://doi.org/10.1016/B978-0-443-21475-2.00010-2

Ghazali R., Mohd Nawi N., Deris M.M., Abawajy J.H., & Arbaiy N. (Ed.). (2022). 5th International Conference on Soft Computing and Data Mining, SCDM 2022. *Lecture Notes in Networks and Systems*, *457 LNNS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130384989&partnerID=40&md5=c92b09bbd4bbc973c3a7a7c5f12e1245

Gupta, N., Thakur, V., Patil, V., Vishnoi, T., & Bhangale, K. (2023). Analysis of Affective Computing for Marathi Corpus using Deep Learning. *Int. Conf. Emerg. Technol., INCET*. 2023 4th International Conference for Emerging Technology, INCET 2023. Scopus. https://doi.org/10.1109/INCET57972.2023.10170346

Harby, F., Alohali, M., Thaljaoui, A., & Talaat, A. S. (2024). Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition. *Computers, Materials and Continua*, *78*(2), 2689–2719. Scopus. https://doi.org/10.32604/cmc.2024.046623

Harika, R., Uday, T., Sirisha, M. L., Sahitya, M. S. L., Drugaanjali, K., & Srinivas, M. S. (2024). A Review of Advancements in Facial Emotion Recognition and Detection Using Deep Learning. *Proc. - Int. Conf. Soc. Sustain. Innov. Technol. Eng., SASI-ITE*, 290–295. Scopus. https://doi.org/10.1109/SASI-ITE58663.2024.00062

Izountar, Y., Benbelkacem, S., Otmane, S., Khababa, A., Zenati, N., & Masmoudi, M. (2021). Towards an adaptive Virtual Reality Serious Game System for Motor Rehabilitation based on Facial Emotion Recognition. *Proc. Int. Conf. Artif. Intell. Cyber Secur. Syst. Priv., AI-CSP*. 2021 Proceedings of the International Conference on Artificial Intelligence for Cyber Security Systems and Privacy, AI-CSP 2021. Scopus. https://doi.org/10.1109/AI-CSP52968.2021.9671149

Jairam, B. G., & Ponnappa, D. (2023). Gesture Based Virtual Assistant For Deaf-Mutes Using Deep Learning Approach. *Int. Conf. Adv. Comput. Commun. Syst., ICACCS*, 1–7. Scopus. https://doi.org/10.1109/ICACCS57279.2023.10112986

Kalra, H. (2023). LSTM Based Feature Learning and CNN Based Classification for Speech Emotion Recognition. *Int. Conf. Data Sci. Netw. Secur., ICDSNS*. 2023 International Conference on Data Science and Network Security, ICDSNS 2023. Scopus. https://doi.org/10.1109/ICDSNS58469.2023.10244802

Kamath, S., Rajendran, R., Wan, Q., Panetta, K., & Agaian, S. S. (2019). TERNet: A deep learning approach for thermal face emotion recognition. Dalam Agaian S.S., Asari V.K., & DelMarco S.P. (Ed.), *Proc SPIE Int Soc Opt Eng* (Vol. 10993). SPIE; Scopus. https://doi.org/10.1117/12.2518708

Keerthana, P. S. M., Vishal, K., Sivani, K. S. S., & Kumar, S. (2022). Covid-19 detection from X-ray scans using Alexa. Dalam Kumar D., Dey T.K., & Dash S. (Ed.), *Int. Conf. Recent Trends Comput. Sci. Technol., ICRTCST - Proc.* (hlm. 121–124). Institute of Electrical and Electronics Engineers Inc.; Scopus. https://doi.org/10.1109/ICRTCST54752.2022.9781958

Kim J., Khan J., Singh M., Tiwary U.S., Sur M., & Singh D. (Ed.). (2022). 13th International Conference on Intelligent Human Computer Interaction, IHCI 2021. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13184 LNCS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127097059&partnerID=40&md5=6951b76fba1842e3e1387894cb0a6b32

Kotwal, R. S., & Gautam, A. (2024). Speech Recognition System based on Wavelet Multi- Resolution Analysis using One-Dimensional CNN-LSTM Network. *Int. Conf. I-SMAC (IoT Soc., Mob., Anal. Cloud), I-SMAC - Proc.*, 882–887. Scopus. https://doi.org/10.1109/I-SMAC61858.2024.10714858

Kreyssig, F. L., & Woodland, P. C. (2020). Cosine-distance virtual adversarial training for semi-supervised speaker-discriminative acoustic embeddings. *Proc. Annu. Conf. Int. Speech. Commun. Assoc., INTERSPEECH*, *2020-October*, 3241–3245. Scopus. https://doi.org/10.21437/Interspeech.2020-2270

Kuś, S., & Szmurło, R. (2021). CNN-based character recognition for a contextless text input system in immersive VR. *CPEE - Int. Conf. "Comput. Probl. Electr. Eng."* CPEE 2021 - 22nd International Conference "Computational Problems of Electrical Engineering." Scopus. https://doi.org/10.1109/CPEE54040.2021.9585252

Kwon, S. (2021). 1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features. *Computers, Materials and Continua*, *67*(3), 4039–4059. Scopus. https://doi.org/10.32604/cmc.2021.015070

Li, Y., Hashim, A. S., Lin, Y., Nohuddin, P. N. E., Venkatachalam, K., & Ahmadian, A. (2024). AI-based visual speech recognition towards realistic avatars and lip-reading applications in the metaverse. *Applied Soft Computing*, *164*. Scopus. https://doi.org/10.1016/j.asoc.2024.111906

Mishra, S., Bhatnagar, N., Prakasam, P., & T. R, S. (2024). Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model. *Multimedia Tools and Applications*, *83*(13), 37603–37620. Scopus. https://doi.org/10.1007/s11042-023-16849-x

Moreno, R. J., Estepa, R. C., & Baquero, J. M. (2022). Audio Commands Recognition Through Deep Learning for Control Mobile Residential Assistant Robot. Dalam Larrondo Petrie M.M., Texier J., Pena A., & Viloria J.A.S. (Ed.), *Proc. LACCEI int. Multi-conf. Eng. Educ. Technol.* (Vol. 2022-July). Latin American and Caribbean Consortium of Engineering Institutions; Scopus. https://doi.org/10.18687/LACCEI2022.1.1.24

Nagar A.K., Jat D.S., Marín-Raventós G., & Mishra D.K. (Ed.). (2022). 5th World Conference on Smart Trends in Systems Security and Sustainability, WS4 2021. *Lecture Notes in Networks and Systems*, *333*. Scopus.

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123277117&partnerID=40&md5=081d0cf7dbfe9ac9fe77a8d358b44b56

Namratha, M., Lokesh, R., Bhat, P., Srikanth, N., & Gagan, M. (2024). InterviewPal-Elevating Interview Automation with Deep Learning and Natural Language Processing Perspectives. *Int. Conf. Emerg. Technol, Comput. Sci. Interdiscip. Appl., ICETCS*. International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications, ICETCS 2024. Scopus. https://doi.org/10.1109/ICETCS61022.2024.10543368

Nayak, S. K., Nayak, A. K., Mishra, S., & Mohanty, P. (2023). Deep Learning Approaches for Speech Command Recognition in a Low Resource KUI Language. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2), 377–386. Scopus.

Pietka E., Badura P., Kawa J., & Wieclawek W. (Ed.). (2019). 7th International Conference on Information Technology in Biomedicine, ITIB 2019. *Advances in Intelligent Systems and Computing*, *1011*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070757059&partnerID=40&md5=9a1378fee4367ce09d2230413cc89c5a

Porwal, A., Tyagi, P. K., & Agarwal, D. K. (2023). Comparative Analysis of Different Neural Network Models for Speaker Gender Recognition by Voice. *Int. Conf. Commun., Secur. Artif. Intell., ICCSAI*, 535–540. Scopus. https://doi.org/10.1109/ICCSAI59793.2023.10421302

Praveen, T. N. V. S., Sivathmika, D., Jahnavi, G., & Bolledu, J. (2023). An In-depth Exploration of ResNet-50 for Complex Emotion Recognition to Unraveling Emotional States. Dalam Kumar R., Kumar R., Gupta M., Gupta M., Srivastava R., & Srivastava R. (Ed.), *Int. Conf. Adv. Comput. Comput. Technol., InCACCT* (hlm. 322–326). Institute of Electrical and Electronics Engineers Inc.; Scopus. https://doi.org/10.1109/InCACCT57535.2023.10141774

Qiao, Z., Zhai, L., Zhang, S., & Zhang, X. (2021). Encrypted 5G Over- The- Top Voice Traffic Identification Based on Deep Learning. *Proc. IEEE Symp. Comput. Commun.*, *2021-September*. Scopus. https://doi.org/10.1109/ISCC53001.2021.9631458

Rajesh Immanuel, R., & Sangeetha, S. K. B. (2023). Decoding Emotions Using Deep Learning Approach to EEG-Based Emotion Recognition. *Intell. Comput. Control Eng. Bus. Syst., ICCEBS*. 2023 Intelligent Computing and Control for Engineering and Business Systems, ICCEBS 2023. Scopus. https://doi.org/10.1109/ICCEBS58601.2023.10449107

Renault E., Boumerdassi S., & Mühlethaler P. (Ed.). (2021). 3rd International Conference on Machine Learning for Networking, MLN 2020. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12629 LNCS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103569737&partnerID=40&md5=358150c0d5bbc44fb1b54c97d3550d7c

Rouhafzay, G., Cretu, A.-M., & Payeur, P. (2021). Transfer of learning from vision to touch: A hybrid deep convolutional neural network for visuo-tactile 3d object recognition. *Sensors (Switzerland)*, *21*(1), 1–15. Scopus. https://doi.org/10.3390/s21010113

Saeed F., Al-Hadhrami T., Mohammed F., & Mohammed E. (Ed.). (2021). 1st International Conference of Advanced Computing and Informatics, ICACIN

2020. *Advances in Intelligent Systems and Computing*, *1188*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096572848&partnerID=40&md5=e632d29e33c0b7360bea80bb6acf3c59

Sain, B., Kumar, R., & Jaiswal, A. (2024). Developmental Sequence in the Comprehension Method of Deep Learning for Classifications of Human Emotions. *Nanotechnology Perceptions*, *20*(S6), 691–702. Scopus. https://doi.org/10.62441/nano-ntp.v20iS6.55

Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*, *8*, 79861–79875. Scopus. https://doi.org/10.1109/ACCESS.2020.2990405

Sangeethapriya, R., & Akilandeswari, J. (2024). Classification of cyberbullying messages using text, image and audio in social networks: A deep learning approach. *Multimedia Tools and Applications*, *83*(1), 2237–2266. Scopus. https://doi.org/10.1007/s11042-023-15538-z

Zaynidinov H., Singh M., Tiwary U.S., & Singh D. (Ed.). (2023). 14th International Conference on Intelligent Human Computer Interaction, IHCI 2022. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13741 LNCS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159493288&partnerID=40&md5=04b170e16d37a23ac220b8e2cce39c07

Zeeshan, M., Qayoom, H., & Hassan, F. (2021). Robust Speech Emotion Recognition System Through Novel ER-CNN and Spectral Features. *Int. Symp. Adv. Electr. Commun. Technol., ISAECT*. 2021 4th International Symposium on Advanced Electrical and Communication Technologies, ISAECT 2021. Scopus. https://doi.org/10.1109/ISAECT53699.2021.9668480

Zhao F. & Miao D. (Ed.). (2024). 1st International Conference on AI-generated Content, AIGC 2023. *Communications in Computer and Information Science*, *1946 CCIS*. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177170160&partnerID=40&md5=9eafc88148762598fada4ad27e9beff2

Zhao, Y., Guo, M., Chen, X., Sun, J., & Qiu, J. (2024). Attention-Based CNN Fusion Model for Emotion Recognition during Walking Using Discrete Wavelet Transform on EEG and Inertial Signals. *Big Data Mining and Analytics*, *7*(1), 188–204. Scopus. https://doi.org/10.26599/BDMA.2023.9020018