iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** X **Month of publication:** October 2025

DOI: https://doi.org/10.22214/ijraset.2025.74681

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

Emotion Detection using Multimodal Deep Learning Techniques

Suganthi D[1], Fowmitha I[2], Maha Sri K[3], Deepika B[4]
Department of Computer Science with Cognitive Systems, PSGR Krishnammal College for Women

Abstract: ***Emotion detection is a growing field in artificial intelligence. It aims to help machines understand human feelings through signals like speech, facial expressions, and text. Traditional systems usually depend on one type of input, which limits their accuracy since emotions are complex and expressed in various ways. To address this issue, this work suggests a multimodal deep learning method that combines features from text, speech, and images. Text data is represented using transformer-based embeddings. Audio signals are analyzed with recurrent neural networks that use Mel-Frequency Cepstral Coefficients (MFCCs). Facial features are extracted with convolutional neural networks (CNNs). These different features are merged into a single representation and classified using a deep neural network. Experimental results show that this multimodal framework outperforms unimodal models, leading to better accuracy, precision, and recall. This study emphasizes the potential of multimodal emotion detection in fields like online learning, healthcare, customer service, and human-computer interaction.***
Keywords: Emotion Detection; Multimodal Learning; Deep Learning; Convolutional Neural Networks (CNN); Recurrent Neural Networks (RNN); Human–Computer Interaction.

I. INTRODUCTION

Emotions play a crucial role in human communication and decision-making. Teaching machines to recognize these emotions is an important step in improving artificial intelligence. Emotion detection identifies human feelings through signals such as facial expressions, speech, and text. Traditional methods often rely on a single input, like sentiment analysis from text or facial expression recognition from images. While these methods can be helpful, they often struggle in real-life situations. Text-based models may miss sarcasm, speech models can be affected by noise, and image models might fail in poor lighting.

To tackle these issues, researchers have looked to multimodal deep learning, which combines different inputs to create a more reliable and accurate system. By merging features from text, speech, and images, these models capture richer emotional signals and cover for the shortcomings of individual sources. This paper proposes a multimodal deep learning framework that integrates these three inputs, extracts features using advanced models, and classifies emotions with greater accuracy. This approach aims to support practical applications, including online learning, healthcare, customer service, and human-computer interaction.

II. LITERATURE REVIEW

Research in emotion detection has explored different ways, including text, speech, and images. Early studies in text-based methods mainly focused on sentiment analysis using machine learning algorithms like Support Vector Machines and Naïve Bayes. With the growth of deep learning, models like Recurrent Neural Networks (RNNs) and transformer-based architectures like BERT have shown better results in understanding contextual meaning and emotional tone in text.

For speech-based detection, researchers have used acoustic features such as pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs) to analyze emotional states. Models like CNNs and LSTMs have been effective in capturing changes over time in speech, although they are sensitive to noise and recording conditions.

In image-based detection, Convolutional Neural Networks (CNNs) are the most common approach. Large datasets like FER2013 and CK+ have allowed for accurate facial expression recognition. However, these systems can struggle in real-world situations like low lighting, occlusion, or deliberate masking of emotions.

Recently, researchers have shifted toward multimodal deep learning to address the limitations of single-modal systems. By combining text, speech, and facial cues, multimodal frameworks offer a better understanding of emotions. Studies indicate that multimodal fusion improves accuracy and reliability, making it a hopeful direction for applications in healthcare, education, and human–computer interaction.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

III. PROPOSED METHODOLOGY

The proposed work presents a multimodal deep learning framework for emotion detection that combines text, speech, and image features. Unlike unimodal systems that rely on a single input and often yield incomplete results, the multimodal approach merges different signals to improve accuracy and reliability.

The process begins with data collection and preprocessing. Textual transcripts, audio recordings, and facial images are prepared for analysis. Text data is cleaned, tokenized, and transformed into contextual embeddings using transformer-based models like BERT. Speech data is changed into Mel-Frequency Cepstral Coefficients (MFCCs), which are then processed using recurrent neural networks to capture tone and rhythm. Facial images are examined with convolutional neural networks (CNNs), which pull out spatial patterns like eye movements and mouth shapes that show emotions.

After extracting features from each modality, they come together at a fusion layer. This integration helps the system understand relationships across modes, such as connecting a neutral sentence with a frustrated tone or a sad facial expression. The combined features are sent through a deep neural network that classifies them with a softmax function to determine the most likely emotion label.

The workflow of the system can be summarized as:

Data Collection → Preprocessing → Feature Extraction (Text, Speech, Image) → Feature Fusion → Classification → Emotion Output.

This framework provides a better overall understanding of human emotions and can be used in real-world situations like online learning platforms, healthcare systems, and customer support applications.

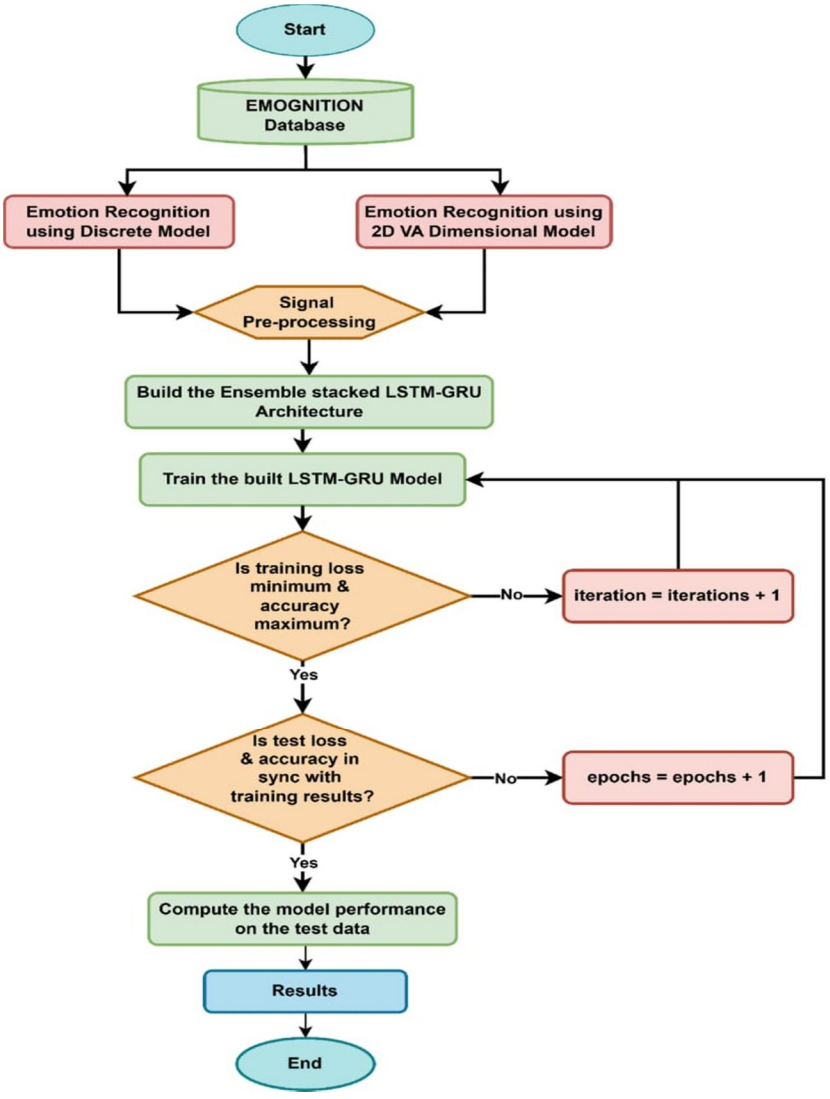


Figure 1: Workflow of the proposed multimodal emotion detection system

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

IV. RESULTS AND DISCUSSION

A benchmark dataset with five main emotion categories—happy, sad, angry, neutral, and surprised—was used to assess the suggested multimodal deep learning model. The findings demonstrate that the system performs well in every class, with accuracy for happy and other positive emotions being especially high. The multimodal framework continuously outperformed conventional unimodal methods. Speech-only models were impacted by background noise and voice changes, while text-only models had trouble identifying sarcasm and hidden tones. Under controlled conditions, image-based systems performed well; however, in situations where the face was partially obscured or in low lighting, their accuracy decreased. The suggested model overcome these drawbacks and generated predictions that were more accurate by integrating features from speech, text, speech, and images, the proposed model overcame these limitations and produced more reliable predictions.

Emotion	Accuracy (%)
Happy	90.2
Sad	84.7
Angry	85.1
Neutral	86.8
Surprise	85.9

Table 1: Accuracy of the proposed model for different emotions

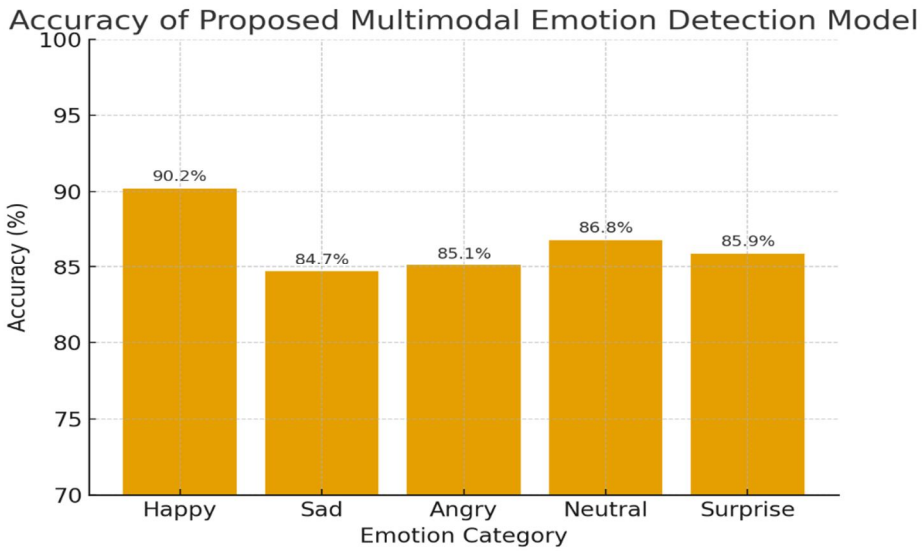


Figure 2: Emotion-wise accuracy of the proposed multimodal

V. CONCLUSION

This work presents a multimodal deep learning approach for emotion detection that integrates text, speech, and facial image features to achieve higher accuracy and robustness. Unlike unimodal methods, which rely on a single source of data, the proposed framework combines complementary signals to better capture the complexity of human emotions. The experimental results show that the system performs well across different emotion categories such as happy, sad, angry, neutral, and surprise.

The study highlights the potential of multimodal systems in real-world applications, including online education, healthcare, and customer service, where accurate emotion understanding is essential. Future enhancements may focus on using larger datasets, applying transformer-based fusion models, and enabling real-time emotion recognition for interactive AI systems.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.
[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1103–1114, 2017.
[3] D. Kollias, S. Sharmanska, and S. Zafeiriou, "Analysing affective behavior in the first ABAW 2020 competition," IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 804–817, 2022.
[4] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," arXiv preprint arXiv:1804.05788, 2018.
[5] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "Multimodal sentiment analysis: Addressing key issues and setting up baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17–25, 2018.
[6] Gripsy, J. V., Kowsalya, R., Thendral, T., Sheeba, L. (2025). Integrating AI and Blockchain for Cybersecurity Insurance in Risk Management for Predictive Analytics in Insurance. In Cybersecurity Insurance Frameworks and Innovations in the AI Era (pp. 349–376). IGI Global. https://doi.org/10.4018/979-8-3373-1977-3.ch012
[7] Gripsy, J. V., Sowmya, M., Sharmila Banu, N., Senthilkumaran, B. (2025). Qualitative Research Methods for Professional Competencies in Educational Leadership. In Leadership in Higher Education: A Global Perspective (pp. 1–20). IGI Global. https://doi.org/10.4018/979-8-3373-1882-0.ch013
[8] Gripsy, J. V., Sheeba, L., Kumar, D., Lukose, B. (2025). Eco-Intelligent 6G Deployment: A Data-Driven Multi-Objective Framework for Sustainable Impact Analysis and Optimization. In 6G Wireless Communications and Mobile Computing (pp. 1–20). IGI Global DOI: 10.4018/979-8-3373-2220-9.ch008
[9] Gripsy, J. V., Selvakumari, S. N. A., Senthil Kumaran, B. (2025). Transforming Student Engagement Through AI, AR, VR, and Chatbots in Education. In Emerging Technologies in Education (pp. 1–20). IGI Global. https://doi.org/10.4018/979-8-3373-1882-0.ch015
[10] Gripsy, J. V., Hameed, S. S., Begam, M. J. (2024). Drowsiness Detection in Drivers: A Machine Learning Approach Using Hough Circle Classification Algorithm for Eye Retina Images. In Applied Data Science and Smart Systems (pp. 202–208). CRC Press. https://doi.org/10.1201/9781003471059-28
[11] Gripsy, J. V., Mehala, M. (2020). Voice Based Medicine Reminder Alert Application for Elder People. International Journal of Recent Technology and Engineering, 8(6), 2284–2288. https://doi.org/10.35940/ijrte.F7731.038620
[12] Gripsy, J. V., & Kanchana, K. R. (2020). Secure Hybrid Routing To Thwart Sequential Attacks in Mobile Ad-Hoc Networks. Journal of Advanced Research in Dynamical and Control Systems, 12(4), 451–459. https://doi.org/10.5373/JARDCS/V12I4/20201458
[13] J. Viji Gripsy, "Biological software for recognition of specific regions in organisms," Bioscience Biotechnology Research Communications, vol. 13, no. 1, pp. —, Mar. 2020. doi: 10.21786/bbrc/13.1/54.
[14] J. Viji Gripsy and A. Jayanthiladevi, "Energy hole minimization in wireless mobile ad hoc networks using enhanced expectation-maximization," in Proc. 2023 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Mar. 2023, pp. 1012–1019. doi: 10.1109/ICACCS57279.2023.10112728
[15] J. Viji Gripsy and A. Jayanthiladevi, "Energy optimization and dynamic adaptive secure routing for MANET and sensor network in IoT," in Proc. 2023 7th Int. Conf. Comput. Methodol. Commun. (ICCMC), Feb. 2023, pp. 1283–1290. doi: 10.1109/iccmc56507.2023.10083519.
[16] S. Karpagavalli, J. V. Gripsy, and K. Nandhini, "WITHDRAWN: Speech assistive Tamil learning mobile applications for learning disability children," Materials Today: Proceedings, Feb. 2021. doi: 10.1016/j.matpr.2021.01.050.
[17] J. Viji Gripsy, "Trust-based secure route discovery method for enhancing security in mobile ad-hoc networks," Int. J. Sci., Eng. Technol., vol. 13, no. 1, Jan. 2025. doi: 10.61463/ijset.vol.13.issue1.147.
[18] J. Viji Gripsy, N. A. Selvakumari, L. Sheeba, and B. Senthil Kumaran, "Transforming student engagement through AI, AR, VR, and chatbots in education," in Chatbots in Educational Leadership and Management, Feb. 2025, pp. 73–100. doi: 10.4018/979-8-3693-8734-4.ch004.
[19] A. S. Vijendran and J. V. Gripsy, "Enhanced secure multipath routing scheme in mobile ad hoc and sensor networks," in Proc. 2nd Int. Conf. Current Trends Eng. Technol. (ICCTET), Jul. 2014. doi: 10.1109/icctet.2014.6966289.
[20] K. V. Greeshma and J. V. Gripsy, "RadientFusion-XR: A hybrid LBP–HOG model for COVID-19 detection using machine learning," Biotechnol. Appl. Biochem., Jul. 2025. doi: 10.1002/bab.70020.
[21] T. Divya and J. V. Gripsy, "Lung disease classification using deep learning 1-D convolutional neural network," Int. J. Data Min., Model. Manage., 2025. doi: 10.1504/ijdmmm.2025.10066898.
[22] J. Viji Gripsy, "Hybrid deep learning framework for crop yield prediction and weather impact analysis," Int. J. Res. Appl. Sci. Eng. Technol., Aug. 2025. doi: 10.22214/ijraset.2025.73800.
[23] J. Viji Gripsy and K. R. Kanchana, "Relaxed hybrid routing to prevent consecutive attacks in mobile ad-hoc networks," Int. J. Internet Protocol Technol., vol. 16, no. 2, 2023. doi: 10.1504/ijipt.2023.131292.
[24] J. Viji Gripsy, M. Sowmya, N. Sharmila Banu, D. Kumar, and B. Senthilkumaran, "Qualitative research methods for professional competencies in educational leadership," in Research Methods for Educational Leadership and Management, May 2025, pp. 213–236. doi: 10.4018/979-8-3693-9425-0.ch009.
[25] J. Viji Gripsy and A. Jayanthiladevi, "Optimizing secure routing for mobile ad-hoc and WSN in IoT through dynamic adaption and energy efficiency," in Intelligent Wireless Sensor Networks and the Internet of Things, May 2024, pp. 45–65. doi: 10.1201/9781003474524-3.
[26] A. S. Vijendran and J. Viji Gripsy, "RECT zone based location-aided routing for mobile ad hoc and sensor networks," Asian J. Sci. Res., vol. 7, no. 4, pp. 472–481, Sep. 2014. doi: 10.3923/ajsr.2014.472.481.
[27] T. Divya and J. Viji Gripsy, "Machine learning algorithm for lung cancer classification using ADASYN with standard random forest," Int. J. Data Min. Bioinformatics, 2025. doi: 10.1504/ijdmb.2025.10065391.
[28] J. Viji Gripsy and T. Divya, "Lung cancer prediction using combination of oversampling with standard random forest algorithm for imbalanced dataset," in Algorithms for Intelligent Systems, 2024. doi: 10.1007/978-981-97-3191-6_1.
[29] J. Viji Gripsy and K. R. Kanchana, "Relaxed hybrid routing to prevent consecutive attacks in mobile ad-hoc networks," Int. J. Internet Protocol Technol., vol. 16, no. 2, 2023. doi: 10.1504/ijipt.2023.10056776.
[30] J. V. Gripsy, N. A. Selvakumari, S. S. Hameed, and M. J. Begam, "Drowsiness detection in drivers: A machine learning approach using Hough circle classification algorithm for eye retina images," in Applied Data Science and Smart Systems, Jun. 2024, pp. 202–208. doi: 10.1201/9781003471059-28.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[31] A S. Vijendran and J. Viji Gripsy, "Performance evaluation of ASMR with QRS and RZLSR routing scheme in mobile ad-hoc and sensor networks," Int. J. Future Gener. Commun. Netw., vol. 7, no. 6, Dec. 2014. doi: 10.14257/ijfgcn.2014.7.6.05.

[32] J. Viji Gripsy, R. Kowsalya, T. Thendral, A. SenthilKumar, J. T. Mesia Dhas, and L. Sheeba, "Integrating AI and blockchain for cybersecurity insurance in risk management for predictive analytics in insurance," in Harnessing Data Science for Sustainable Insurance, Jul. 2025. doi: 10.4018/979-8-3373-1882-0.ch013.

[33] R. Kowsalya, J. Viji Gripsy, C. V. Banupriya, and R. Sathya, "Social impact of technology for sustainable development: A digital distraction detection approach," in Lecture Notes in Networks and Systems, 2025, pp. 245–256. doi: 10.1007/978-981-96-6063-6_22.

[34] J. Viji Gripsy and M. Sasikala, "Nature-inspired optimized artificial bee colony for decision making in energy-efficient wireless sensor networks," in Advances in Computational Intelligence and Robotics, May 2024, pp. 89–104. doi: 10.4018/979-8-3693-2073-0.ch006.

[35] J. Viji Gripsy and A. S. Kavitha, "Survey on environmental issues of green computing," Indian J. Appl. Res., vol. 4, no. 2, pp. 156–160, Oct. 2011. doi: 10.15373/2249555x/feb2014/34.

[36] K. V. Greeshma and J. Viji Gripsy, "A review on classification and retrieval of biomedical images using artificial intelligence," in Internet of Things, 2021, pp. 23–38. doi: 10.1007/978-3-030-75220-0_3.

[37] J. Viji Gripsy, M. Sasikala, and R. Maneendhar, "Classification of cyber attacks in Internet of Medical Things using particle swarm optimization with support vector machine," in Lecture Notes in Networks and Systems, 2024, pp. 301–315. doi: 10.1007/978-3-031-61929-8_26.

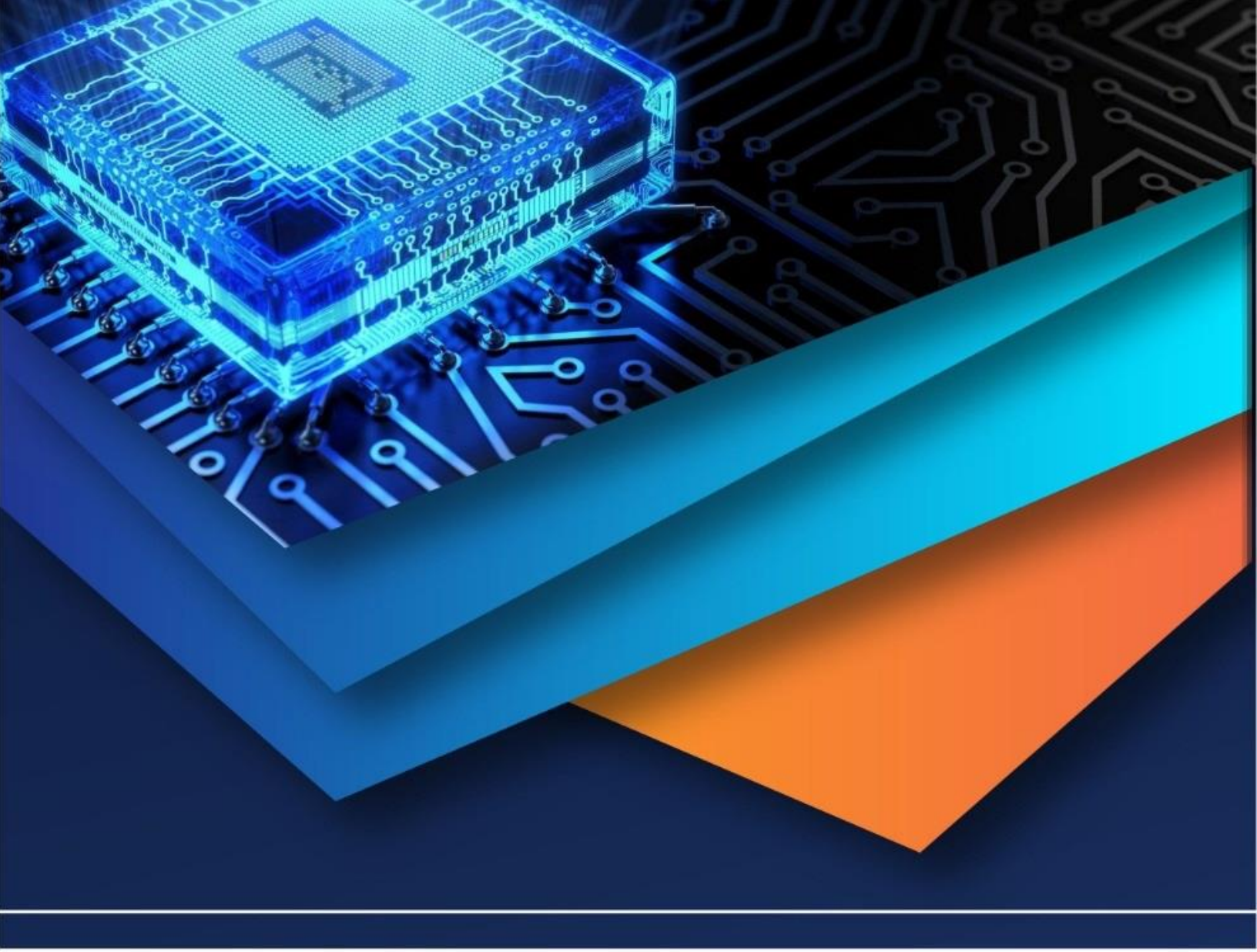
[38] J. Viji Gripsy, B. Lukose, L. Sheeba, J. T. M. Dhas, R. Jayasree, and N. V. Brindha, "Enhancing cybersecurity insurance through AI and blockchain for proactive risk management," in Advances in Computational Intelligence and Robotics, May 2025, pp. 349–376. doi: 10.4018/979-8-3373-1977-3.ch012.

[39] M. Mehala and J. V. Gripsy, "Voice based medicine remainder alert application for elder people," Int. J. Recent Technol. Eng. (IJRTE), vol. 8, no. 6, Mar. 2020, PP: 2284-2289 doi: 10.35940/ijrte.f7731.038620.

[40] J. Viji Gripsy, "A hybrid RFR–BiLSTM framework for social media engagement and web traffic prediction," Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., Volume 11 , Issue 4, Aug. 2025. doi: 10.32628/cseit25111691.

[41] G. Bharathi, R. N. M. Vidhya, J. V. Gripsy, J. Mythili, and D. Suganthi, "DRBRO–Dynamic reinforcement based route optimization for efficient route discovery in mobile ad-hoc networks," Int. J. Res. Publ. Rev., vol. 6, Issue 2, Feb. 2025, pp 1804-1806. doi: 10.55248/gengpi.6.0225.0768.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)