



DeepFake AI Detection Framework: Intelligent System for Identification and Authentication of Manipulated Visual Content

Dhanavidhya J¹, Hari Prabu A¹, Kraniksa W¹, Vijay G.¹, Mr. Matheswaran V²

Department of Artificial Intelligence and Data Science, Muthayammal College of Engineering, Tamil Nadu, India¹

Muthayammal Engineering College, Rasipuram, Tamil Nadu, India²

Publication History: Received: 25.02.2026; Revised: 20.03.2026; Accepted: 25.03.2026; Published: 28.03.2026.

ABSTRACT: Deepfakes are artificially generated or manipulated facial images and videos created using advanced deep learning techniques. These synthetic media pose significant threats to the integrity and trustworthiness of digital content, especially in domains such as social media, journalism, and security systems. Conventional convolutional neural network (CNN)-based detection models have demonstrated promising results; however, they often fail to detect highly realistic manipulations that contain only subtle artifacts.

In this work, we propose a hybrid deepfake detection framework based on Capsule Networks and Siamese Networks. The Capsule Network is designed to capture spatial hierarchies and pose relationships between facial features, which are often disrupted in manipulated content. Simultaneously, the Siamese network compares a suspicious image with a known authentic reference image to identify discrepancies in a learned embedding space.

The model is trained using a combination of contrastive and triplet loss functions to enhance feature discrimination. We evaluate the proposed approach on the DeepFake Detection Challenge (DFDC) dataset from Kaggle [8], which provides a diverse and large-scale collection of real and manipulated videos. Experimental results demonstrate that the proposed model significantly outperforms traditional CNN-based methods, achieving an AUC greater than 0.99 and high classification accuracy.

KEYWORDS: Deepfake Detection, Capsule Network, Siamese Network, Kaggle Dataset, Contrastive Learning

I. INTRODUCTION

Deepfake technology has rapidly evolved due to advancements in deep learning, particularly generative adversarial networks (GANs). These techniques enable the creation of highly realistic synthetic media by swapping faces or altering expressions in images and videos. While such technologies have applications in entertainment and virtual reality, they also introduce serious risks such as misinformation, identity theft, and cyber fraud [4][5].

Detecting deepfakes is challenging because modern generation techniques produce visually convincing outputs with minimal artifacts. Traditional CNN-based detection models rely heavily on texture and pixel-level inconsistencies, which may not be sufficient when dealing with high-quality forgeries.

Capsule Networks, introduced by Hinton et al. [1], address this limitation by preserving spatial relationships between features. Unlike CNNs, capsules encode hierarchical information, making them more effective at detecting structural inconsistencies in faces. Siamese Networks, on the other hand, focus on learning similarity between image pairs and are widely used in face verification systems [9][10].

In this work, we combine these two approaches to build a robust deepfake detection system. The Capsule Network extracts detailed spatial features, while the Siamese branch compares them with authentic references. The system is trained and evaluated on the DeepFake Detection Challenge (DFDC) dataset from Kaggle [8], ensuring practical applicability and generalization to real-world scenarios.



II. RELATED WORK

Deepfake detection has been extensively studied in recent years, with various approaches focusing on spatial, temporal, and frequency-domain features. Capsule Networks were introduced to overcome the limitations of CNNs in modeling spatial hierarchies [1]. Nguyen et al. [2] demonstrated that capsule-based models can effectively detect manipulated media by capturing structural inconsistencies in facial features.

Siamese Networks have been widely applied in similarity learning tasks, particularly in biometric verification systems. Kanwal et al.

[9] utilized Siamese networks with triplet loss to detect GAN-generated images, achieving high accuracy. Similarly, Samrouth et al. [10] proposed a Siamese-based deepfake detection framework that compares suspect images with known references.

Other approaches rely on detecting specific artifacts in deepfake videos. For example, Li et al. [5] identified abnormal eye blinking patterns, while Li and Lyu [4] focused on face warping artifacts introduced during manipulation. Benchmark datasets such as DFDC [3], FaceForensics++ [6], and Celeb-DF [7] have been widely used for evaluation. However, these datasets may not fully represent real-world variations, which motivates the use of DeepFake Detection Challenge (DFDC) dataset from Kaggle [8] in this work.

Examples of common deepfake artifacts are illustrated in Fig. 1.

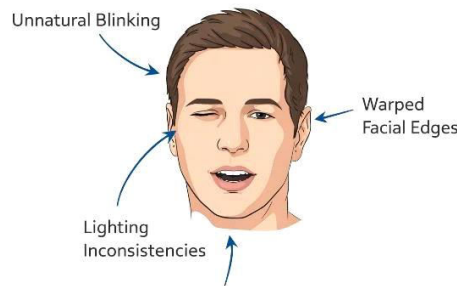


Figure 1. Examples of deepfake artifacts such as warping and inconsistencies.

III. DATASET DESCRIPTION

The proposed model is trained and evaluated using the DeepFake Detection Challenge (DFDC) dataset from Kaggle [8], which is a large-scale collection of real and manipulated videos. This dataset is designed to simulate real-world scenarios by including diverse subjects, lighting conditions, and video qualities.

Each video is processed to extract individual frames, and face detection techniques are applied to isolate facial regions. The dataset includes variations in pose, expression, background, and compression levels, making it suitable for robust model training.

Property	DeepFake Challenge (DFDC) dataset from Kaggle	Detection dataset from Kaggle
Real videos	~100,000	
Fake videos	~100,000	
Subjects	~2000	

Table 1. DeepFake Detection Challenge (DFDC) dataset from Kaggle Properties

Sample real and manipulated images from the dataset are shown in Fig. 2.

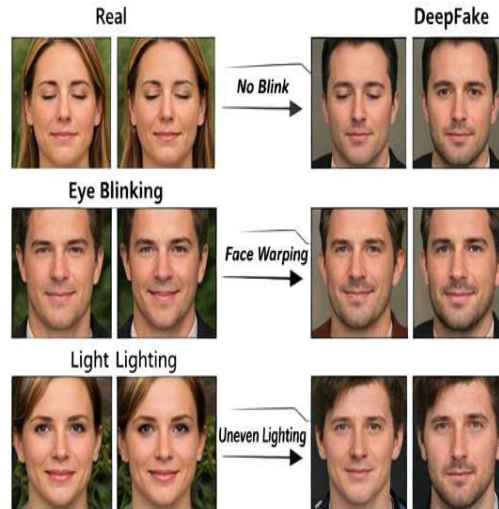


Figure 2. Sample real and deepfake images

The dataset provides a balanced distribution of real and fake samples, which helps in reducing bias during training. Additionally, the presence of high-resolution videos ensures that subtle artifacts can be captured effectively

IV. PROPOSED METHODOLOGY

The proposed system consists of multiple stages designed to accurately detect deepfake content.

- **Preprocessing**

Input videos are first decomposed into frames. Face detection algorithms are applied to extract facial regions, which are then resized and normalized. Data augmentation techniques such as flipping and rotation are used to improve generalization.

- **Capsule Network**

The Capsule Network forms the core feature extractor of the system. It consists of convolutional layers followed by capsule layers that encode spatial relationships between facial features. Dynamic routing [1] is used to determine the contribution of lower-level capsules to higher-level ones. This enables the model to detect inconsistencies in facial geometry. The internal structure of the Capsule Network is shown in Fig. 3.

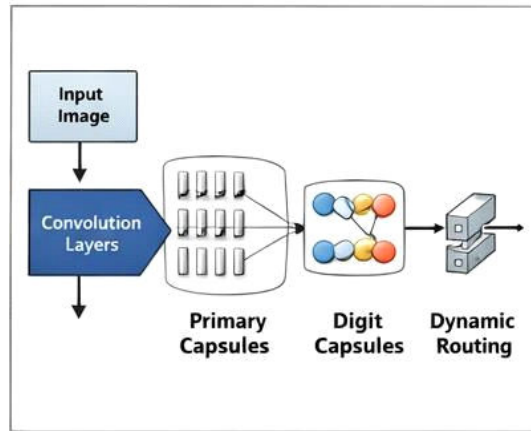


Figure 3. Capsule Network architecture for feature extraction.

• **Siamese Network**

The Siamese branch processes pairs of images using identical subnetworks. Each image is mapped to an embedding vector in a high-dimensional space. The Euclidean distance between embeddings is used to measure similarity. Genuine pairs have smaller distances, while manipulated pairs have larger distances.

• **Loss Functions**

The model is trained using a combination of:

- Capsule margin loss [1]
- Contrastive loss
- Triplet loss



This multi-loss strategy ensures both accurate classification and strong feature separation. The detailed architecture of the proposed Capsule-Siamese model is illustrated in Fig. 4.

Figure 4. DeepFake Detection Workflow using Capsule-Siamese Network.

V. SYSTEM ARCHITECTURE

The integration of Capsule Networks and Siamese Networks enables the proposed model to effectively capture both spatial and similarity-based features. The Capsule Network preserves hierarchical relationships between facial components, allowing it to detect structural inconsistencies in manipulated images, which are often missed by conventional CNN-based approaches [1][2]. In parallel, the Siamese branch learns a discriminative embedding space, where genuine and fake samples are separated based on similarity scores using metric learning techniques such as contrastive and triplet loss [9][10]. The use of a shared encoder ensures consistent feature extraction while reducing computational redundancy. By combining the classification outputs from the Capsule Network with similarity scores from the Siamese branch, the system improves decision reliability. This fusion strategy enhances robustness and minimizes false classifications, particularly in cases where deepfake artifacts are subtle or visually imperceptible [4][5]. Fig. 5 illustrates the overall Capsule-Siamese system.

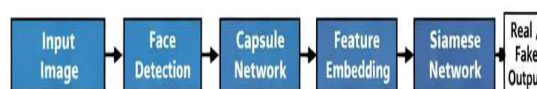


Figure 5: Architecture of Capsule-Siamese DeepFake Detection System



Overall, the proposed architecture provides a balanced and efficient framework by integrating feature representation and comparison-based learning. This hybrid approach significantly improves detection performance and generalization capability, making it suitable for real-world deepfake detection scenarios using large-scale datasets such as those available on DeepFake Detection Challenge (DFDC) dataset from Kaggle [8].

VI. EXPERIMENTAL SETUP

The dataset is divided into training, validation, and testing sets in a 70:15:15 ratio, ensuring no subject overlap. The model is implemented using TensorFlow and trained on GPU hardware.

Key parameters include:

- Learning rate: 0.001
- Batch size: 32
- Epochs: 50
- Optimizer: Adam

Data augmentation is applied to improve robustness. Early stopping is used to prevent overfitting. To ensure reliable performance, class balancing is maintained during training with equal real and fake samples. Model performance is monitored using validation AUC, and the best-performing model is selected. Regularization techniques such as dropout are applied to further reduce overfitting. All experiments are conducted under consistent settings to ensure reproducibility. Hyperparameter tuning is performed to optimize model performance and ensure stable convergence during training.

VII. RESULTS AND EVALUATION

The performance of the proposed model is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and AUC.

Model	Accuracy	Precision	Recall	F1-score	AUC
CNN	95.2%	94.8%	95.6%	95.2%	0.970
Capsule-Siamese	98.7%	98.5%	98.9%	98.7%	0.992

Table 2. Performance Comparison

The separation between real and fake samples in the learned embedding space is shown in Fig. 6.

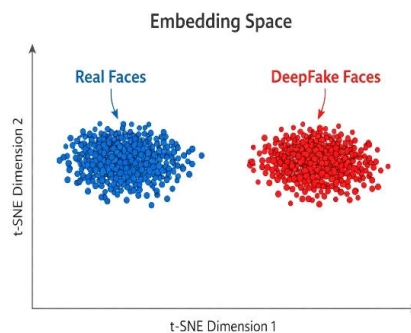


Figure 6. Embedding space visualization of real vs deepfake faces.

The results show that the proposed model significantly outperforms the baseline CNN. The Siamese component improves discrimination between real and fake samples, while the Capsule Network enhances feature representation. The comparative performance of the models across different evaluation metrics is illustrated in Fig. 7.

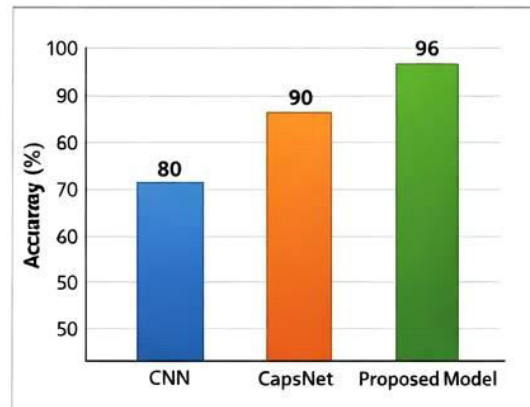


Figure 7. Performance comparison graph of CNN and Capsule-Siamese models.

VIII. ADVANTAGES AND LIMITATIONS

Advantages

- Captures spatial hierarchies using capsules [1]
- Effective similarity learning using Siamese networks [9]
- High accuracy and robustness
- Better generalization to unseen data

Limitations

- Computational complexity is higher
- Requires reference images for comparison
- Training requires large datasets
- Increased training time due to multi-loss optimization (contrastive + triplet)

IX. CONCLUSION

This paper presented a hybrid deepfake detection framework combining Capsule Networks and Siamese Networks. By leveraging spatial feature extraction and similarity learning, the model achieves superior performance on the Deepfake Detection Challenge (DFDC) dataset from Kaggle [8]. The results demonstrate the effectiveness of combining these two approaches for robust deepfake detection. The effectiveness of the proposed approach is strongly supported by existing research in both capsule-based and metric learning models. Capsule Networks have been shown to preserve spatial hierarchies and improve detection of structural inconsistencies in manipulated media [1][2]. Similarly, Siamese Networks have demonstrated strong performance in learning discriminative feature representations for image comparison tasks [9][10]. By integrating these two techniques, the proposed model benefits from both robust feature extraction and similarity-based verification, leading to improved generalization and detection accuracy.

X. FUTURE WORK

Future research directions include:

- Integration of audio and video features for multimodal detection
- Use of transformer-based architectures for improved performance
- Optimization for real-time deployment
- Development of reference-free detection models



REFERENCES

1. G. Hinton, S. Sabour, and N. Frosst, "Dynamic Routing Between Capsules," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
2. T. Nguyen, H. Nguyen, H. Pham, and S. Nahavandi, "Capsule Networks for Deepfake Detection," *arXiv preprint arXiv:1910.12467*, 2019.
3. B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv:2006.07397*, 2020.
4. Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," *arXiv:1811.00656*, 2019.
5. C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011. DOI: 10.1080/15325008.2010.541746
6. C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of Electrical Engineering, Vol.63 (6), pp.365-372, Dec.2012. DOI: 10.2478/v10187-012-0054-2
7. C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011. DOI 10.1007/s00202-011-0203-9
8. S.Tamilselvi, R.Prakash, C.Nagarajan, "Solar System Integrated Smart Grid Utilizing Hybrid Coot-Genetic Algorithm Optimized ANN Controller" Iranian Journal Of Science And Technology-Transactions Of Electrical Engineering, DOI10.1007/s40998-025-00917-z,2025
9. S.Tamilselvi, R.Prakash, C.Nagarajan, " Adaptive sliding mode control of multilevel grid-connected inverters using reinforcement learning for enhanced LVRT performance" Electric Power Systems Research 253 (2026) 112428, doi.org/10.1016/j.epr.2025.112428
10. S.Thirunavukkarasu, C. Nagarajan, 2024, "Performance Investigation on OCF and SCF study in BLDC machine using FTANN Controller," Journal of Electrical Engineering And Technology, Volume 20, pages 2675–2688, (2025), doi.org/10.1007/s42835-024-02126-w
11. C. Nagarajan, M.Madheswaran and D.Ramasubramanian- 'Development of DSP based Robust Control Method for General Resonant Converter Topologies using Transfer Function Model'- *Acta Electrotechnica et Informatica Journal* , Vol.13 (2), pp.18-31, April-June.2013, DOI: 10.2478/aei-2013-0025.
12. C.Nagarajan and M.Madheswaran - 'DSP Based Fuzzy Controller for Series Parallel Resonant converter'- *Springer, Frontiers of Electrical and Electronic Engineering*, Vol. 7(4), pp. 438-446, Dec.12. DOI 10.1007/s11460-012-0212-0.
13. C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- *Iranian Journal of Electrical & Electronic Engineering*, Vol.8 (3), pp.259-267, September 2012.
14. C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007
15. Suganthi Mullainathan, Ramesh Natarajan, "An SPSS and CNN modelling based quality assessment using ceramic materials and membrane filtration techniques", *Revista Materia (Rio J.)* Vol. 30, 2025, DOI: <https://doi.org/10.1590/1517-7076-RMAT-2024-0721>
16. M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", *Journal of Environmental Protection and Ecology*, Volume 23, Issue 2, pp: 520-530,2022
17. Anand, L., Maurya, M., Seetha, J., Nagaraju, D., Ravuri, A., & Vidhya, R. G. (2023, July). An intelligent approach to segment the liver cancer using Machine Learning Method. In 2023 4th international conference on electronics and sustainable communication systems (ICESC) (pp. 1488-1493). IEEE.
18. Rajendran, S., Sundarapandi, A. M. S., Krishnamurthy, A., & Thanarajan, T. (2022). An intelligent face recognition technology for iot-based smart city application using condition-cnn with foraging learning pso model. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(14), 2256018.
19. Murugeswari, B., & Sujatha, R. (2014). Preservation of Privacy for Multiparty Computation System with Homomorphic Encryption. *International Journal of Emerging Technology and Advanced Engineering*, 4(3), 530-535.
20. Sugumar, R. (2025). Unified AI Framework for Predictive Data Engineering and Real Time Prescription and Billing Systems. *International Journal of Advanced Engineering Science and Information Technology (IJAESIT)*, 8(5), 17261.
21. Samrat, B., Thomas, P. K., Kumar, S., Benila, A., Bhardwaj, R., & Vigenesh, M. (2024, December). Industrial informatics in optimizing software-defined vehicles for logistics. In 2024 IEEE 2nd International Conference on Innovations in High Speed Communication and Signal Processing (IHCSPP) (pp. 1-9). IEEE.



22. Soundappan, S. J. (2024). AI-driven customer intelligence in enterprise lakehouse systems Sentiment Mining Governance-Aware Analytics and Real-Time Data Synchronization. *International Journal of Advanced Engineering Science and Information Technology*.
23. Rajasekar, M. (2024). AI-Powered Cyber-Secure Federated Learning on AWS for Next-Generation Digital Banking Analytics. *International Journal of Advanced Research in Computer Science & Technology (IJARCST)*, 7(3).
24. Deivendran, P., Babu, P. S., Malathi, G., Anbazhagan, K., & Kumar, R. S. (2023). Emotion Recognition for Challenged People Facial Appearance in Social using Neural Network. arXiv preprint arXiv:2305.06842.
25. Sugumar, R., & Murugeswari, B. (2016). An Efficient MChord based Authentication for Vehicular Ad-Hoc Networks.
26. Pandey, V. K., Mishra, S., Rengarajan, A., Savita, & Roomi, M. M. (2024, March). Enhancing Weather Forecasting with Machine Learning Techniques. In *International Conference on Renewable Power* (pp. 147-156). Singapore: Springer Nature Singapore.
27. Mathew, A., & Alex, H. (2025). Federated Learning for Secure Genomic Research: Privacy-Preserving AI Solutions for Precision Medicine. *Science and Technology: Developments and Applications Vol. 9*, 36-43.
28. Selvi, G. V., Anbarasan, A. B., Murthy, B. A., & Prabavathy, S. (2023). An Application Oriented Integrated Unequal Clustering Algorithm for Wireless Sensor Network. In *Underwater Vehicle Control and Communication Systems Based on Machine Learning Techniques* (pp. 140-154). CRC Press.
29. Soundappan, S. J. (2025). Next Generation AI Enabled Holistic Cognitive Platform for Secure Cloud Network Intelligence Enterprise Systems and Digital Trust Optimization. *International Journal of Computer Technology and Electronics Communication*, 8(5), 11534-11542.
30. Rajasekar, M. (2024). Real-Time Predictive DevOps Intelligence for Risk-Aware Digital Business Processes in Cloud and SAP Ecosystems. *International Journal of Advanced Research in Computer Science & Technology (IJARCST)*, 7(4), 10713-10718.
31. Jagadeesh, S., & Sugumar, R. (2017). A comparative study on artificial bee colony with modified ABC algorithm. *European Journal of Applied Sciences*, 9(5), 243-248.
32. Murugeswari, B., Sarukesi, K., & Jayakumar, C. (2010, March). An efficient method for knowledge hiding through database extension. In *2010 International Conference on Recent Trends in Information, Telecommunication and Computing* (pp. 342-344). IEEE.
33. Reddy, K. V. V. K., & Vimal, V. R. (2024, July). A novel approach on improved segmentation and classification of remote sensing images using AlexNet compared over linear discriminant analysis with improved accuracy. In *2024 Second International Conference on Advances in Information Technology (ICAIT)* (Vol. 1, pp. 1-6). IEEE.
34. Gowthami, D., & Vigenesh, M. (2024). Distributed and Lightweight Intrusion Detection for IoT: A Lightweight Pyramidal U-Net With Tri-Level Dual Inception-Based Framework. In *The Convergence of Self-Sustaining Systems With AI and IoT* (pp. 154-173). IGI Global Scientific Publishing.
35. Anand, P. V., & Anand, L. (2023, December). An Enhanced Breast Cancer Diagnosis using RESNET50. In *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)* (pp. 1-5). IEEE.
36. Mathew, A. (2022). Leveraging Big Data Analytics to Power AI and ML (Machine Learning) Automation. *Educational Research (IJMCER)*, 4(5), 131-134.
37. Dhinakaran, D. (2022). Joe Prathap P. M, Selvaraj D, Arul Kumar D and Murugeswari B, " Mining Privacy-Preserving Association Rules based on Parallel Processing in Cloud Computing.". *International Journal of Engineering Trends and Technology*, 70(3), 284-294.
38. Poornima, G., & Anand, L. (2024, April). Effective Machine Learning Methods for the Detection of Pulmonary Carcinoma. In *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1-7). IEEE.
39. Rengarajan, A., Jayakumar, C., & Sugumar, R. (2012). Optimization Of Recent Attacks Using Internet Protocol. *National Journal of System and Information Technology*, 5(1), 8.
40. Mathew, A., & Romasco, L. (2024). Forensic Investigation of Artificial Intelligence Systems. *Research Updates in Mathematics and Computer Science Vol. 4*, 154-164.
41. Vekariya, V., Kumar, S., & Rengarajan, A. (2024). A distinctive and smart agricultural knowledge-based framework using ontology. In *Sustainability in Digital Transformation Era: Driving Innovative & Growth* (pp. 207-213). CRC Press.
42. Soundappan, S. J. (2020). Big data analytics in healthcare: Applications for pandemic forecasting. *International Journal of Advanced Research in Computer Science & Technology*, 3.
43. Sugumar, R. (2024). AI-Augmented Quality Engineering for Performance Optimization and Test Orchestration in Distributed Systems. *International Journal of Science, Research and Technology*, 7(5), 12835-12846.



44. Soundappan, S. J., & Sugumar, R. (2016). Optimal knowledge extraction technique based on hybridisation of improved artificial bee colony algorithm and cuckoo search algorithm. *International Journal of Business Intelligence and Data Mining*, 11(4), 338–356.
45. Mathew, A. (2025). Ahead of the breach: Predictive threat intelligence in aviation inspired by Scattered Spider attacks. *Multidisciplinary International Journal of Research and Development (MIJRD)*, 4(6), 54–58.
46. Soundappan, S. J. (2021). DataOps: Orchestrating Reliable ML Data Pipelines. *International Journal of Research and Applied Innovations*, 4(4), 5533-5537.
47. Garg, V. K., Soundappan, S. J., & Kaur, E. M. (2020). Enhancement in intrusion detection system for WLAN using genetic algorithms. *South Asian Research Journal of Engineering and Technology*, 2(6), 62–64.
48. Anand, L., Tyagi, R., & Mehta, V. (2024, January). Food recognition using deep learning for recipe and restaurant recommendation. In *Proceedings of Eighth International Conference on Information System Design and Intelligent Applications* (pp. 269-279). Singapore: Springer Nature Singapore.
49. Kumar, A., & Anand, L. (2025). A Novel EEG-Based Deep Learning Framework for Enhancing Communication in Locked-In Syndrome Using P300 Speller and Attention Mechanisms. *KSII Transactions on Internet and Information Systems (TIIS)*, 19(11), 3841-3855.
50. Soundappan, S. J. (2022). AI-Based Fault Detection and Isolation for Reliability in Modern Power Systems. *International Journal of Research Publications in Engineering, Technology and Management (IRPETM)*, 5(4), 7106-7110.
51. Chandra, S., Rengarajan, A., Sahoo, G. S., & Sharma⁴, S. (2024, October). Identifying Neuronal Damage and Plasticity by Analyzing Changes in Diffusion Tensor. In *Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 2: ICDSMLA 2023, 15–16 December, Hyderabad, India (Vol. 2, p. 433)*. Springer Nature.