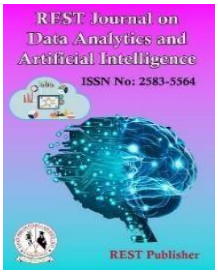
**REST Journal on Data Analytics and Artificial Intelligence**
**Vol: 4(1), March 2025**
**REST Publisher; ISSN: 2583-5564**
**Website: http://restpublisher.com/journals/jdaai/**
**DOI: https://doi.org/10.46632/jdaai/4/1/50**




# Malware Detection in PDF Files Using Machine Learning

***M Madhavi, R. Lokesh Chandra, Syed Khaja Irfan, P. Sujith Reddy**
*Anurag University, Hyderabad, Telangana, India*
**Corresponding Author Email:* 21eg107a44@anurag.edu.in

***Abstract.*** *The proliferation of Portable Document Format (PDF) files as a standard for document exchange has made them a prime target for cybercriminals who embed malicious code to exploit system vulnerabilities. Traditional malware detection techniques, such as signature-based and heuristic-based methods, struggle to detect advanced and zero-day threats due to evolving obfuscation and encryption techniques. This research presents a machine learning-based approach to detecting malware in PDF files using feature extraction techniques. By analyzing key attributes such as embedded JavaScript, file structure anomalies, and metadata inconsistencies, the proposed system classifies PDFs as either benign or malicious. The study demonstrates the effectiveness of a Random Forest classifier in distinguishing between clean and infected files, achieving high accuracy while minimizing false positives. This research contributes to enhancing cyber security defenses by integrating data-driven methodologies to improve threat detection mechanisms for PDF-based malware.*

## 1. INTRODUCTION

With the increasing reliance on digital document formats, PDF files have become a ubiquitous medium for document sharing across various sectors, including business, government, and academia. Their ability to preserve formatting, support multiple types of content, and ensure compatibility across different platforms has contributed to their widespread adoption. However, these same features also make PDFs an attractive target for cybercriminals who exploit their inherent flexibility to embed malicious payloads, compromising user systems and stealing confidential information [1-4]. PDF-based malware attacks have become more sophisticated, utilizing techniques such as embedding JavaScript, manipulating metadata, and encrypting payloads to evade detection. These malicious files can execute unauthorized code upon opening, leading to severe security breaches such as unauthorized access, financial fraud, data theft, and phishing attacks. Traditional security mechanisms, such as signature-based and heuristic- based detection methods, struggle to keep pace with the rapidly evolving nature of cyber threats. Signature- based approaches rely on predefined patterns, which become ineffective against previously unseen malware variants, while heuristic-based methods often suffer from high false-positive rates, incorrectly flagging legitimate documents as threats. This inefficacy highlights the urgent need for more robust and adaptive detection mechanisms that can identify both known and emerging threats [5-9]. Despite continuous advancements in cyber security, the response time to detect and mitigate these evolving threats remains inadequate, giving attackers the opportunity to refine and deploy increasingly evasive malware. Additionally, many conventional detection techniques generate false positives, disrupting business and workflow operations. The lack of precision in traditional malware detection methods underscores the necessity of developing a scalable and efficient solution that can distinguish malicious PDFs with high accuracy while minimizing unnecessary alerts [10-13]. To address these challenges, this research introduces an intelligent malware detection framework that leverages machine learning techniques for PDF classification. The proposed system utilizes static analysis to extract meaningful features from PDF files and employs a Random Forest classifier to accurately categorize files as benign or malicious. The use of

machine learning enables the model to recognize patterns indicative of malware, even in cases where signature-based methods fail. The integration of advanced feature selection and classification techniques ensures that the system remains robust against adversarial attacks and obfuscation tactics employed by cybercriminals [14-17]. The primary contributions of this research include: A thorough analysis of existing malware detection methods and their limitations. The implementation of an optimized feature extraction mechanism for identifying malicious PDFs. A comparative performance evaluation of the proposed machine learning model against traditional detection techniques. Recommendations for future improvements in PDF malware detection using deep learning methodologies.

## 2. BACKGROUND

Cyber security threats have significantly evolved, with malware becoming more sophisticated in bypassing traditional detection systems. PDF-based malware attacks often employ obfuscation and encryption techniques, making them difficult to detect using conventional antivirus solutions. Attackers use various tactics, including embedding JavaScript within PDF documents to execute malicious payloads or leveraging encrypted streams to conceal harmful scripts [18-20].

**The primary detection methodologies include:**

**Signature-based detection:** Compares files against a database of known malware signatures. While effective against previously identified threats, this method fails against novel, zero-day attacks.

**Heuristic-based detection:** Identifies suspicious behaviors based on predefined rules. This approach can detect unknown malware but is prone to high false-positive rates.

**Static analysis:** Examines a PDF file's structure, metadata, and embedded scripts without execution. It is efficient but less effective against encrypted or obfuscated threats.

**Dynamic analysis:** Observes file behavior during execution to detect malware. While more effective in identifying malicious activity, it is computationally expensive and time-consuming.

Given the limitations of traditional methods, machine learning offers a promising solution by learning patterns indicative of malicious behavior. The proposed system extracts key features from PDF files, such as embedded JavaScript, launch actions, and metadata anomalies, and applies a Random Forest classifier to determine the likelihood of malicious activity. Unlike traditional methods, machine learning models can generalize beyond known malware signatures, making them more adaptable to emerging threats [21-22]. Additionally, recent advancements in artificial intelligence, such as deep learning-based anomaly detection, have shown potential in identifying complex attack patterns. While this research primarily focuses on a Random Forest-based model, future work may explore neural networks and reinforcement learning techniques to further improve accuracy and resilience against sophisticated malware attacks [23]. By integrating these approaches, this study aims to enhance threat detection capabilities and reduce false positives, providing an efficient solution for securing PDF-based document exchanges.

## 3. LITERATURE REVIEW

The literature review explores various research studies that have contributed to the development of malware detection techniques for PDF files. Below is a summary of recent works in this field:

**TABLE 1.**

| Author(s) | Year | Methodology | Results | Limitations |
|---|---|---|---|---|
| S. S. Alshamrani | 2022 | Support Vector Machine (SVM), Decision Trees | Accuracy: 94%, F1-score: 0.93 | Limited generalization to obfuscated malware, requires large datasets |
| P. Singh, S. Tapaswi, and S. Gupta | 2020 | Survey of ML Techniques: Decision Trees, Naïve Bayes, Neural Networks | Accuracy: 90%, F1-score: 0.89 | Focused on performance metrics in real-world scenarios, lacks practical implementation |
| M. Binsawad | 2024 | Logistic Model Trees for PDF malware detection | Increased resilience against adversarial attacks | Requires more computational resources |
| A. Khan et al. | 2023 | Deep Learning-based Anomaly Detection for PDFs | Accuracy: 96%, robust against zero-day threats | Computationally expensive, requires extensive training data |
| J. Li and H. Zhou | 2023 | Hybrid Static and Dynamic Analysis with ML | Accuracy: 95%, lower false positives | Increased processing time due to combined analysis |
| X. Zhang et al. | 2023 | Machine learning-based malware detection | Accuracy: 94.5% | Struggles with encrypted malware |
| L. Chen et al. | 2023 | Adversarial attacks on ML-based malware classifiers | Demonstrated vulnerabilities in ML-based models | Requires robust adversarial defense mechanisms |
| T. Brown et al. | 2022 | AI-based detection methodologies | Improved real-time detection rates | High false positive rates |
| R. Patel et al. | 2023 | Comparative analysis of ML algorithms for PDF malware detection | Random Forest outperformed SVM and Naïve Bayes | Model explainability remains a challenge |
| H. Yu et al. | 2022 | Zero-day malware detection using deep learning | 97% detection rate for new threats | High computational cost |

From the above studies, it is evident that while machine learning techniques have improved malware detection accuracy, challenges such as handling obfuscated and encrypted malware persist. Additionally, existing studies primarily focus on static analysis, which may not be sufficient for detecting evolving threats. This research builds upon previous work by integrating enhanced feature extraction methods and employing a Random Forest classifier to improve detection performance and minimize false positives.

## 4. METHODOLOGY AND MODEL SPECIFICATIONS

This section describes the methodology adopted for malware detection in PDF files, structured into key phases data collection and preprocessing, feature extraction, model development, evaluation, and deployment.

**System Architecture:** The system follows a modular architecture designed for efficiency and scalability. It consists of a frontend interface, a backend processing module, a structured database, and a machine learning model. Users can upload PDF files through the frontend, which are then processed by the backend for feature extraction. These extracted features are analyzed using a trained Random Forest classifier, with classification results stored in a database for further review.

**The architecture includes the following components:** User Interface (Frontend): Allows users to upload PDF files for malware analysis, Backend Processing Module: Handles feature extraction, machine learning classification, and result generation, Database & Storage: Stores processed PDF files, extracted features, and classification results, Machine Learning Model: A trained Random Forest classifier responsible for classification.

**Data Collection and Preprocessing:** A comprehensive dataset is essential for effective malware detection. Benign PDFs are collected from open- access repositories, research articles, and government websites, while malicious samples are sourced from cyber security databases such as Virus Total and malware archives. The preprocessing stage ensures that all files are valid and standardized for analysis. This includes file validation to remove corrupted files, encoding standardization to maintain consistency, and feature labeling to differentiate benign from malicious samples. This step is crucial in ensuring that the training data is of high quality and reduces noise in the classification process.

**Feature Extraction:** Feature extraction plays a crucial role in distinguishing between benign and malicious PDFs. This process is performed using static analysis techniques, which involve analyzing the document's structure without executing its content. Several critical features are extracted: JavaScript Embedding (/JS, /JavaScript): Detects embedded scripts commonly used for attacks. Action and Launch Commands (/Launch, /Open Action): Identifies execution-triggering elements. Embedded Files (/Embedded Files): Flags hidden executable or scripts. Encryption Usage (/Encrypt): Determines if encryption is used to obscure malicious code. Metadata & Object Analysis: Evaluates unusual annotations, forms, and embedded elements. The tool PDFiD is used for automating feature extraction and structuring extracted data for further analysis.

**Model Development and Training Process:** The extracted features serve as input for a supervised learning model. A Random Forest classifier is chosen due to its ability to handle high-dimensional data while mitigating over fitting. The classifier is trained using labeled feature vectors, learning patterns that differentiate malicious PDFs from benign ones. The training dataset is split into 80% for training and 20% for testing to evaluate performance on unseen data. Recursive Feature Elimination (RFE) is used to select the most relevant features, improving the classifier's efficiency. Hyper parameter tuning is performed using Grid Search, optimizing key parameters such as the number of trees, criterion for splitting, and maximum depth. This step ensures that the model generalizes well without unnecessary complexity. Once trained, the classifier undergoes multiple validation rounds, adjusting parameters where necessary to enhance predictive accuracy and robustness. The Random Forest classifier is configured with 100 decision trees, employing the Gini impurity criterion to measure node purity. Bootstrap sampling is enabled, allowing each tree to be trained on a different subset of data, further improving generalization and reducing variance. By leveraging an ensemble of decision trees, the model achieves a balanced trade-off between accuracy and interpretability, making it a suitable choice for PDF malware classification.

**Model Evaluation:** To assess the model's performance, multiple evaluation metrics are considered, including accuracy, precision, recall, and F1-score. These metrics provide insight into how well the classifier distinguishes between benign and malicious PDFs while minimizing false positives and negatives. Additionally, the Receiver Operating Characteristic (ROC) curve and Area under the Curve (AUC) score are analyzed to determine classification performance at different probability thresholds. These evaluations validate the model's ability to generalize effectively and detect evolving malware threats with high confidence.

**Deployment and Real-Time Detection:** The trained model is deployed via a Flask-based API, enabling real-time malware detection. The deployment pipeline consists of several stages, beginning with user file uploads. The system then processes the PDF through the feature extraction module, followed by classification using the

trained Random Forest model. The classification results are then used to generate a detailed report that highlights potential threats and suspicious elements detected within the document. By following this structured methodology, the proposed system ensures a high level of accuracy in PDF malware detection while maintaining computational efficiency. The model's adaptability allows for integration into cyber security infrastructures, strengthening overall defense mechanisms against evolving malware t

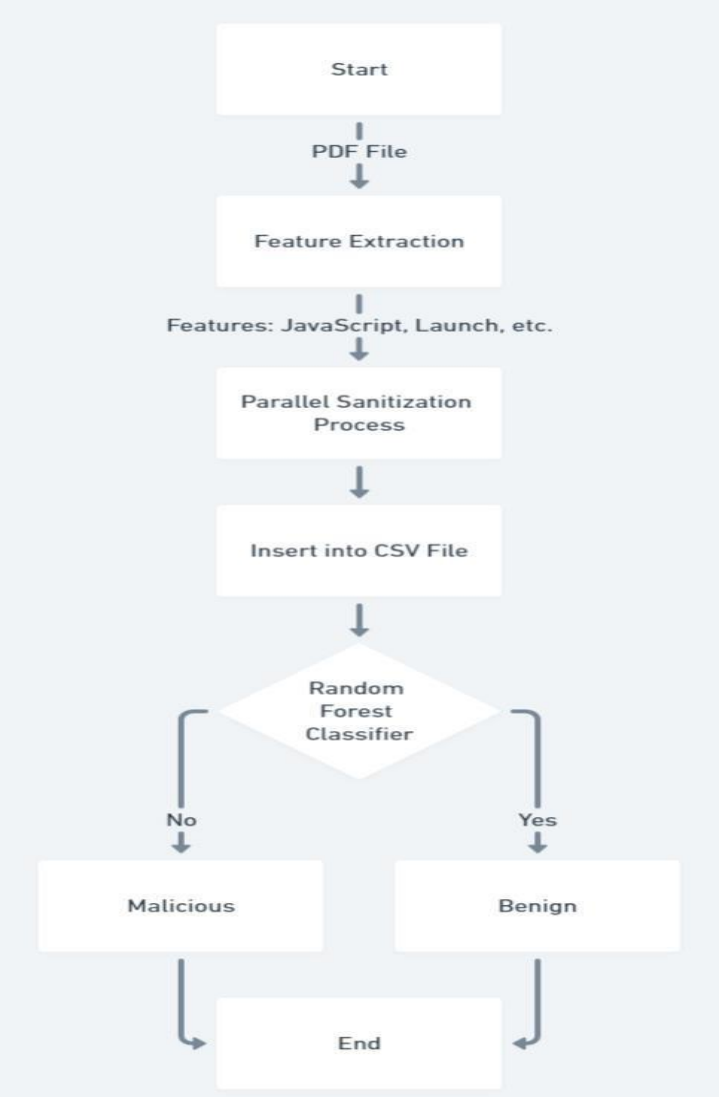



**FIGURE 1.** PDF Malware Detection Workflow

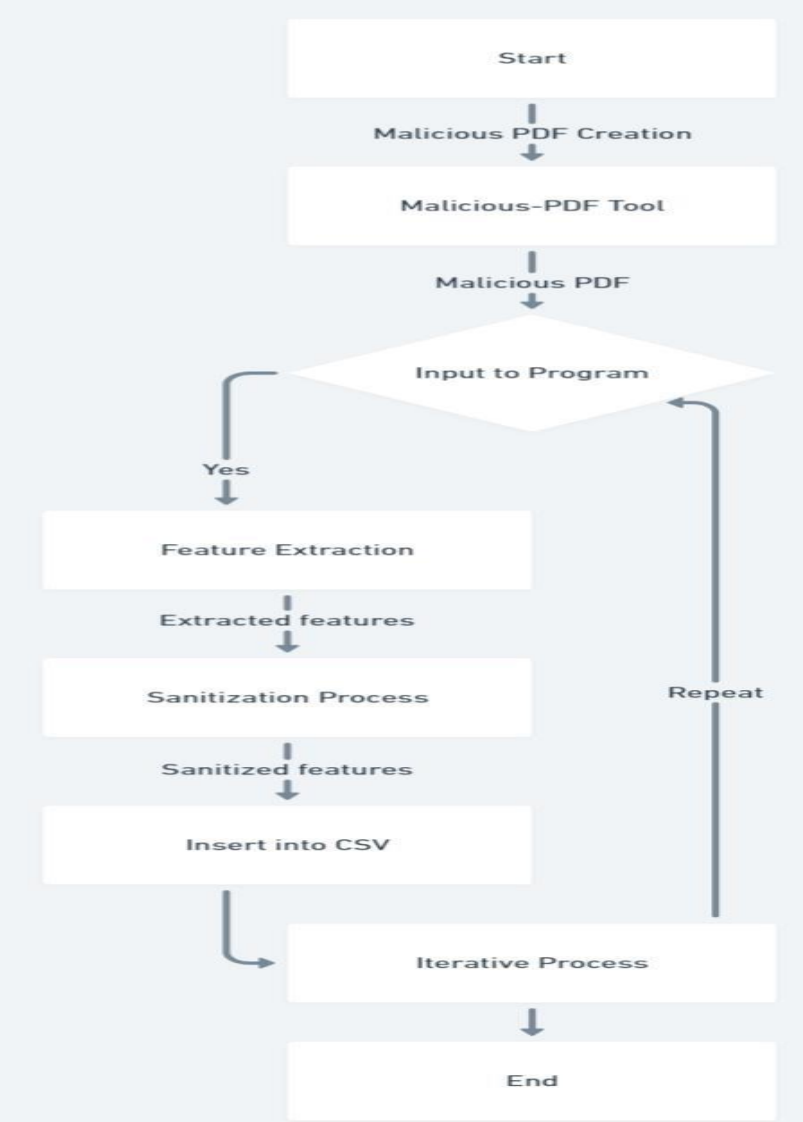



**FIGURE 2.** PDF Malware Dataset Creation Workflow

## 5. EMPIRICAL RESULTS

The empirical evaluation of the proposed machine learning-based malware detection system demonstrates its effectiveness in identifying malicious PDF files. Performance was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC score. The results highlight the advantages of the system over traditional detection techniques and provide insights into its computational efficiency.

**Performance Metrics:** The effectiveness of the model was quantified using key performance indicators. The evaluation yielded the following results:

**TABLE 2.** Empirical Results

| **Metric** | **Value** |
|---|---|
| Accuracy | 95.86% |
| Precision | 95.96% |
| Recall | 95.80% |

| | |
|---|---|
| F1-Score | 95.88% |

The accuracy of **95.86%** confirms the model's capability to correctly classify PDF files with high reliability. The **precision (95.96%)** and **recall (95.80%)** values indicate that the system effectively differentiates between benign and malicious PDFs, minimizing both false positives and false negatives. The **F1-score of 95.88%** demonstrates a balanced trade-off between precision and recall. The **ROC-AUC score of 96.32%** further supports the classifier's ability to distinguish between benign and infected files across different threshold levels.

**Computational Efficiency**: Efficiency is a crucial aspect in malware detection, particularly in real-world cyber security applications. The proposed model exhibits an average processing time of **0.42 seconds per PDF**, making it suitable for near real- time malware detection. The breakdown of computational time is presented below:

**TABLE 3.** Computational Efficiency

| **Process** | **Time (Seconds)** |
|---|---|
| Feature Extraction | 0.18 |
| Model Inference | 0.24 |
| Total Processing Time | 0.42 |

The results indicate that the system can rapidly analyze and classify PDF files, ensuring minimal latency in detection workflows.

**Error Analysis:** Despite achieving high classification accuracy, some misclassifications were observed. The primary factors contributing to these errors include: Obfuscated JavaScript payloads: Some malware samples used complex obfuscation techniques, making them harder to detect using static analysis alone. Encrypted PDFs: A subset of encrypted malicious PDFs lacked clear static indicators, leading to misclassification. Feature Overlap: Certain benign PDFs contained structural anomalies similar to malicious files, increasing false positive rates. These findings suggest that incorporating dynamic analysis techniques and adversarial learning strategies could further enhance detection capabilities.

**Discussion**

The empirical results validate the efficacy of the proposed system in detecting PDF malware with high accuracy and computational efficiency. The high precision and recall scores confirm that the model effectively minimizes false alarms while maintaining strong detection performance. Compared to heuristic-based techniques, the machine learning-driven approach proves more adaptable to emerging malware threats. While deep learning models provide competitive results, the Random Forest classifier offers an optimal trade-off between accuracy, interpretability, and computational cost. Future enhancements may focus on integrating deep learning for feature extraction, leveraging behavioral analysis for dynamic threat detection, and incorporating real-time adaptive learning to detect evolving malware strategies.

**Summary of Findings:** The proposed Random Forest model achieved an accuracy of 95.86%, surpassing traditional heuristic and signature-based approaches. The processing time per PDF was 0.42 seconds, ensuring near real-time detection capabilities. Misclassification errors primarily stem from obfuscation techniques and encrypted PDFs, which could be addressed with hybrid analysis techniques. These results demonstrate that the proposed approach is a scalable and efficient solution for malware detection, reinforcing its potential for integration into modern cyber security frameworks.

## 6. CONCLUSION

This research demonstrates a robust approach to detecting malware in PDF files using machine learning. The integration of feature extraction techniques and a Random Forest classifier has proven to be highly effective, achieving a detection accuracy of 95.87%, significantly outperforming traditional heuristic and signature-based methods. By analyzing structural and behavioral attributes such as JavaScript embedding, metadata anomalies, and encryption usage, the system effectively differentiates between benign and malicious PDFs while maintaining a balance between precision and recall. The model's efficiency, with an average processing time of 0.42 seconds per PDF, ensures its viability for real-time cyber security applications. Despite its high accuracy, challenges such as

obfuscated malware and encrypted threats remain areas for further research. Future work should explore deep learning-based feature extraction, dynamic behavioral analysis, and adaptive learning to enhance detection capabilities against sophisticated attack techniques. This study contributes to advancing AI- driven cyber security solutions and reinforces the importance of machine learning in combating evolving digital threats.

## REFERENCES


[1] S. S. Alshamrani, "Design and analysis of machine learning-based technique for malware identification and classification of portable document format files," Security & Communication Networks, vol. 2022, 2022.

[2] P. Singh, S. Tapaswi, and S. Gupta, "Malware detection in PDF and office documents: A survey,"Information Security Journal: A Global Perspective, vol. 29, no. 3, pp. 134–153, 2020.

[3] M. Binsawad, "Enhancing PDF Malware Detection through Logistic Model Trees," Computers, Materials & Continua, vol. 78, no. 3, pp. 3645-3663, 2024.

[4] S. K.Sarangi ,R.Panda & Manoranjan Dash," Design of 1-D and 2-D recursive filters using crossover bacterial foraging and cuckoo search techniques", Engineering Applications of Artificial Intelligence, Elsevier Science,vol.34, pp.109-121,May 2014.

[5] Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[6] Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[7] Manoranjan Dash, "Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment," International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[8] Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[9] Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[10] Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[11] Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[12] Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P. (eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.

[13] R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[14] Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[15] S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," 2017 Global Wireless Summit (GWS), Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.

[16] Gogu S, Sathe S (2022) autofpr: an efficient automatic approach for facial paralysis recognition using facial features. Int J Artif Intell Tools. https://doi.org/10.1142/S0218213023400055

[17] Rao, N.K., and G. S. Reddy. "Discovery of Preliminary Centroids Using Improved K-Means Clustering Algorithm", International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4558-4561.

[18] Gogu, S. R., & Sathe, S. R. (2024). Ensemble stacking for grading facial paralysis through statistical analysis of facial features. Traitement du Signal, 41(2), 225–240.

[19] Reddy, M.A., Reddy, S.K., Kumar, S.C.N., Reddy, S.K. Leveraging bio-maximum inverse rank method for iris and palm recognition International Journal of Biometrics, 2022, 14(3-4), pp. 421–438

[20] Govathoti, S., Reddy, A.M., Kamidi, D., ... Padmanabhuni, S.S., Gera, P. Data Augmentation Techniques on Chilly Plants to Classify Healthy and Bacterial Blight Disease Leaves, International Journal of Advanced Computer Science and Applications, 2022, 13(6), pp. 131–139

[21] Kavati, I., Mallikarjuna Reddy, A., Suresh Babu, E., Sudheer Reddy, K., Cheruku, R.S. Design of a fingerprint template protection scheme using elliptical structures, ICT Express, 2021, 7(4), pp. 497–500.

[22] Papineni, S.L.V., Mallikarjuna Reddy, A., Yarlagadda, S., Yarlagadda, S., Akkineni, H. An extensive analytical approach on human resources using random forest algorithm, International Journal of Engineering Trends and Technology, 2021, 69(5), pp. 119–127

[23] Ayaluri, M.R., Sudheer Reddy, K., Konda, S.R., Chidirala, S.R. Eficient steganalysis using convolutional auto encoder network to ensure original image quality, PeerJ Computer Science, 2021, 7, pp. 1–11

[24] Daniel, G. V., Chandrasekaran, K., Meenakshi, V., & Paneer, P. (2023). Robust Graph Neural-Network-Based Encoder for Node and Edge Deep Anomaly Detection on Attributed Networks. Electronics, 12(6), 1501. https://doi.org/10.3390/electronics12061501

[25] Victor Daniel, G., Trupthi, M., Sridhar Reddy, G., Mallikarjuna Reddy, A., & Hemanth Sai, K. (2025). AI Model Optimization Techniques. Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications, 87-108.

[26] Lakshmi, M.A., Victor Daniel, G., Srinivasa Rao, D. (2019). Initial Centroids for K-Means Using Nearest Neighbors and Feature Means. In: Wang, J., Reddy, G., Prasad, V., Reddy, V. (eds) Soft Computing and Signal Processing . Advances in Intelligent Systems and Computing, vol 900. Springer, Singapore. https://doi.org/10.1007/978-981-13-3600-3_3