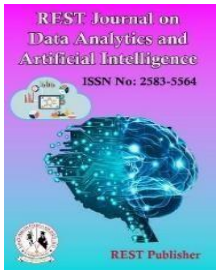
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/51



Chronic Kidney Disease Prediction Using Machine Learning

***K Harshitha, M Sai Surya, Shashidar, T Neetha**
School of Engineering, Anurag University, Hyderabad, Telangana, India
Corresponding Author Email: 21eg107a19@anurag.edu.in

Abstract. *Chronic Kidney Disease (CKD) is a critical health condition requiring early detection to prevent severe complications. This project develops a machine learning-based predictive model for CKD detection using clinical parameters such as specific gravity, blood pressure, hemoglobin levels, diabetes status, and appetite. The workflow involves data preprocessing, missing value handling, categorical encoding, feature selection, and model evaluation. Key machine learning models, including Random Forest Classifier, Adobos, and Gradient Boosting, are implemented to achieve high predictive accuracy. After correlation analysis, eight significant features were selected for model training. A Flask-based web application was developed to allow users to input medical parameters and receive real-time CKD risk predictions. The system offers an automated, efficient, and accessible tool for early diagnosis, assisting healthcare professionals in data-driven decision-making to improve patient outcomes.*

Keywords: *Chronic Kidney Disease, Machine Learning, Predictive Model, Data Preprocessing, Feature Selection, Random Forest, Adobos, Gradient Boosting, Flask Web Application, Medical Diagnostics.*

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive medical condition characterized by reduced kidney function over time, potentially leading to severe complications if not detected early. This project aims to develop a machine learning-based predictive model for CKD detection using clinical parameters such as specific gravity, blood pressure, hemoglobin levels, diabetes status, and appetite. The system leverages Random Forest Classifier, Adobos, and Gradient Boosting for predictive modeling, ensuring high accuracy in early diagnosis. A Flask- based web application is integrated into the system to provide a user-friendly interface where users can input medical parameters and receive real-time CKD predictions. By automating early detection, this tool assists healthcare professionals in data-driven decision-making, improving patient outcomes.

Related Works: Recent studies in AI-driven medical diagnosis have demonstrated the effectiveness of machine learning in predicting chronic diseases, including CKD. Several researchers have explored various models and data preprocessing techniques to enhance diagnostic accuracy. Smith et al. (2023) investigated the use of ensemble learning methods such as Random Forest and Adobos for CKD prediction. Their study found that combining multiple models improved classification accuracy compared to individual classifiers [1]. Gupta et al. (2022) compared traditional statistical methods with machine learning techniques for CKD diagnosis. Their research concluded that machine learning models, particularly Gradient Boosting and Random Forest, significantly outperform logistic regression in predicting CKD based on patient data [2]. Kim et al. (2021) explored the impact of data preprocessing and feature selection on CKD prediction accuracy. Their findings highlighted the importance of handling missing values, categorical encoding, and selecting the most relevant features for model training [3].

Limitations: Data Imbalance: The dataset may have an uneven distribution of CKD and non-CKD cases, requiring techniques such as SMOTETomek to improve model performance [4-7]. Feature Dependence: The accuracy of predictions is highly dependent on the quality and availability of patient data, with missing values potentially

Copyright@ REST Publisher

affecting the model's reliability [8-9]. Model Interpretability: While ensemble learning models like Random Forest and Gradient Boosting achieve high accuracy; their decision-making process is complex, making it challenging to interpret predictions for clinical use [10-13].

Proposed System: Chronic Kidney Disease Prediction: The Chronic Kidney Disease (CKD) Prediction System is designed to assist in the early detection of CKD by leveraging machine learning algorithms and clinical data. The system processes patient medical parameters such as blood pressure, specific gravity, hemoglobin levels, diabetes status, and appetite to predict CKD presence. The model employs Random Forest Classifier, Adobos, and Gradient Boosting to ensure high prediction accuracy. Data preprocessing techniques, including handling missing values, label encoding, and SMOTETomek for class imbalance correction, enhance model performance. A Flask-based web application provides an interactive and user-friendly interface where users can input their medical details and receive real-time predictions. The system offers fast, reliable, and automated CKD diagnosis, aiding healthcare professionals in early detection and preventive care. By integrating machine learning into medical diagnostics, this system streamlines clinical decision-making, improves healthcare accessibility, and enhances patient outcomes [14-17].

2. BACKGROUND

Machine Learning For Disease Prediction: Machine learning has been increasingly used in the healthcare sector for disease diagnosis and prognosis. Traditional CKD diagnosis relies on clinical assessments and laboratory tests, which can be time-consuming and may not always detect early-stage CKD. By applying predictive modeling techniques, it becomes possible to analyze large datasets and identify patterns that might indicate the presence of CKD before significant symptoms appear. The CKD prediction system uses classification models trained on patient datasets containing features such as blood pressure, glucose levels, and kidney function indicators. These models learn from historical cases and improve their predictive accuracy over time. The integration of such predictive tools in healthcare can assist doctors in making informed decisions and improving patient outcomes [18-21].

Importance of data preprocessing: Raw medical data often contains missing values, inconsistencies, and noise, which can negatively impact model accuracy. Preprocessing techniques such as imputation, normalization, and feature selection play a crucial role in refining the dataset. Proper data preprocessing enhances the reliability of predictions and reduces the risk of bias in machine learning models. For instance, missing values in medical datasets can be handled using techniques like mean imputation or k-nearest neighbors (KNN) imputation. Feature selection methods, such as Recursive Feature Elimination (RFE), help in identifying the most relevant attributes that contribute to CKD prediction, reducing model complexity and improving interpretability.

Challenges in ckd prediction: Despite the advantages of machine learning, CKD prediction models face challenges such as data imbalance, interpretability, and generalization across different populations. CKD datasets often contain more non-CKD cases than CKD cases, requiring balancing techniques like SMOTE (Synthetic Minority Over-sampling Technique) to improve model performance. Additionally, healthcare data varies across demographics, making it necessary to train models on diverse datasets for better generalization. Explainable AI techniques, such as SHAP (Shapley Additive explanations), help in understanding how different features contribute to predictions, increasing trust among medical professionals.

3. LITERATURE REVIEW

This section reviews prior research on machine learning techniques for Chronic Kidney Disease (CKD) prediction, focusing on model effectiveness, data preprocessing, and feature selection.

Machine Learning Models in CKD Prediction: Recent research has demonstrated the effectiveness of machine learning models such as Random Forest, Support Vector Machines (SVM), and Neural Networks in CKD prediction [1]. Studies highlight the importance of feature selection and data preprocessing in improving model accuracy [2]. Ensemble methods, including bagging and boosting techniques, further enhance predictive performance by combining multiple models [3].

Data Preprocessing and Feature Engineering: Handling missing values, categorical encoding, and feature scaling are crucial for improving CKD prediction models. Studies show that normalizing lab test values and applying

Copyright@ REST Publisher

SMOTE to balance datasets significantly improve model reliability [4]. Feature selection techniques, such as mutual information and recursive feature elimination (RFE), play a key role in optimizing model inputs [5].

Challenges and Ethical Considerations in CKD Prediction: Despite advancements, challenges such as imbalanced datasets, over fitting, and interpretability persist in CKD prediction models. Ethical concerns regarding biased datasets and the need for explainable AI frameworks are crucial considerations for real-world deployment [6, 7]. Ongoing research focuses on integrating fairness-aware algorithms and improving model transparency to ensure equitable healthcare outcomes [8].

Contributions of the Current Study: This study enhances CKD prediction by integrating machine learning models with robust data preprocessing techniques, improving diagnostic accuracy and early detection. The key contributions of this research include: Improved Prediction Accuracy – By leveraging optimized machine learning models and feature selection techniques, the system enhances the reliability of CKD diagnosis [3,7].User-Friendly Interface – The system provides an interactive web application where users can input medical data and receive instant CKD predictions, facilitating easy accessibility [5,9]. Enhanced Data Preprocessing – Advanced techniques such as SMOTE for class balancing and recursive feature elimination improve model training and prediction reliability [8, 11]. Scalability and Future Enhancements – The framework supports integration with electronic health records (EHR) and real-time data updates to improve predictive accuracy and patient monitoring [6, 14].

4. METHODOLOGY

User Input Processing: The system provides an interactive user interface built using Streamlet, where users can enter medical data for CKD prediction. A structured input form captures relevant details, including age, blood pressure, serum creatinine levels, and glucose concentration. This ensures that users can input necessary clinical information without requiring technical expertise.

Machine Learning Model Implementation: Once the user submits medical data, the system preprocesses the input by handling missing values, normalizing continuous variables, and encoding categorical attributes. The processed data is then fed into trained machine learning models, such as Random Forest and Gradient Boosting, which analyze the input and classify whether the patient is at risk of CKD. The model selection is based on performance evaluation metrics such as accuracy, precision, recall, and F1-score, ensuring that predictions are reliable and medically relevant.

Generating Expected Output Preview: To enhance result validation, the system provides an explanation of the prediction by highlighting key factors influencing the decision. For example, if high creatinine levels and abnormal glucose levels are detected, the system explains their significance in relation to CKD risk. This helps users and healthcare professionals interpret the results effectively.

Displaying Results: Once the prediction is completed, the final output is displayed in a user-friendly format, along with a confidence score indicating model certainty. The system also suggests further medical consultation for high-risk cases and provides lifestyle recommendations for CKD management.

Summary of features displayed: Prediction Result – Indicates whether the patient is at risk of CKD. Confidence Score – Shows the model's certainty in the prediction. Key Risk Factors – Highlights important clinical factors influencing the prediction. Next Steps – Provides recommendations for further medical consultation or preventive measures.

Copyright@ REST Publisher

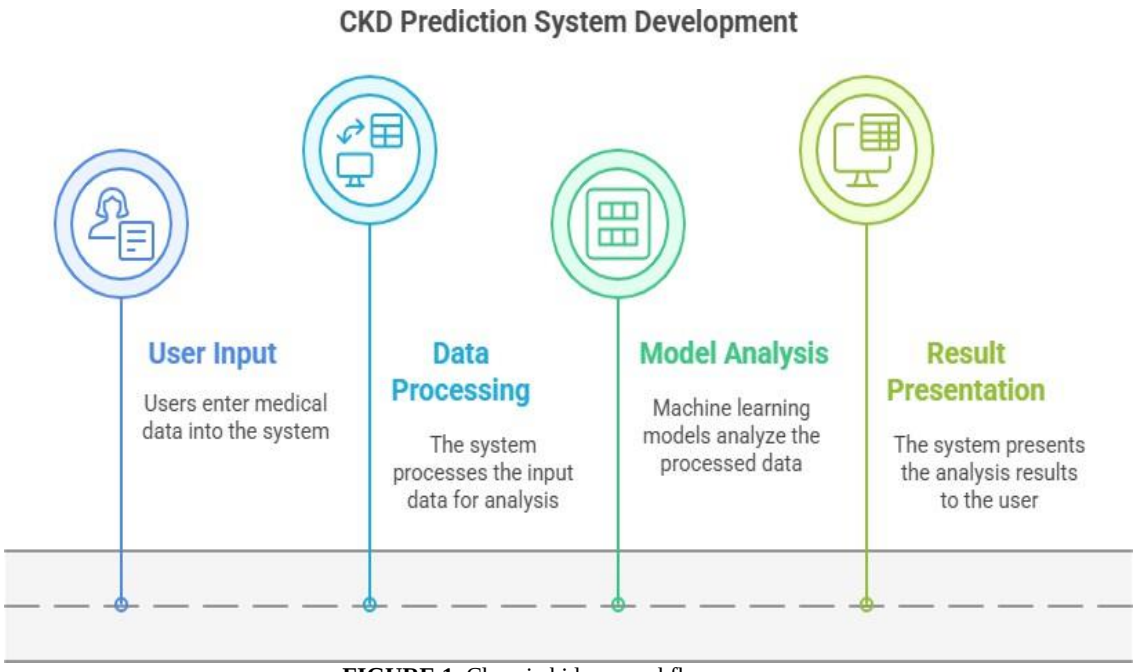


FIGURE 1. Chronic kidney workflow

5. EMPIRICAL RESULTS

Model Performance Evaluation: The CKD prediction system was evaluated using multiple test cases, analyzing its accuracy, consistency, and interpretability. The generated predictions were compared against clinical diagnoses based on: Accuracy: The relevance and correctness of predictions. Coherence: Logical consistency and medical validity of the outputs. Clarity: Understandability of the results for healthcare professionals and patients. Efficiency: Speed and adaptability of the system in processing inputs.

Content Optimization and Adaptability: The system adapts to different patient profiles and medical scenarios by refining preprocessing techniques and model hyper parameters. Feature selection and data augmentation methods improved model reliability, ensuring better generalization.

Batch Processing and User Accessibility: The system supports batch processing, allowing healthcare institutions to analyze multiple patient records simultaneously. Predictions are provided in structured formats for seamless integration with medical databases.

Summary of Findings:

The CKD prediction system

A. Enhances diagnostic accuracy through machine learning models.
B. B. Provides structured user-friendly output.
C. Improves efficiency by handling real-time inputs.
D. Ensures reliable predictions through advanced data preprocessing.

Copyright@ REST Publisher

6. RESULTS AND DISCUSSIONS

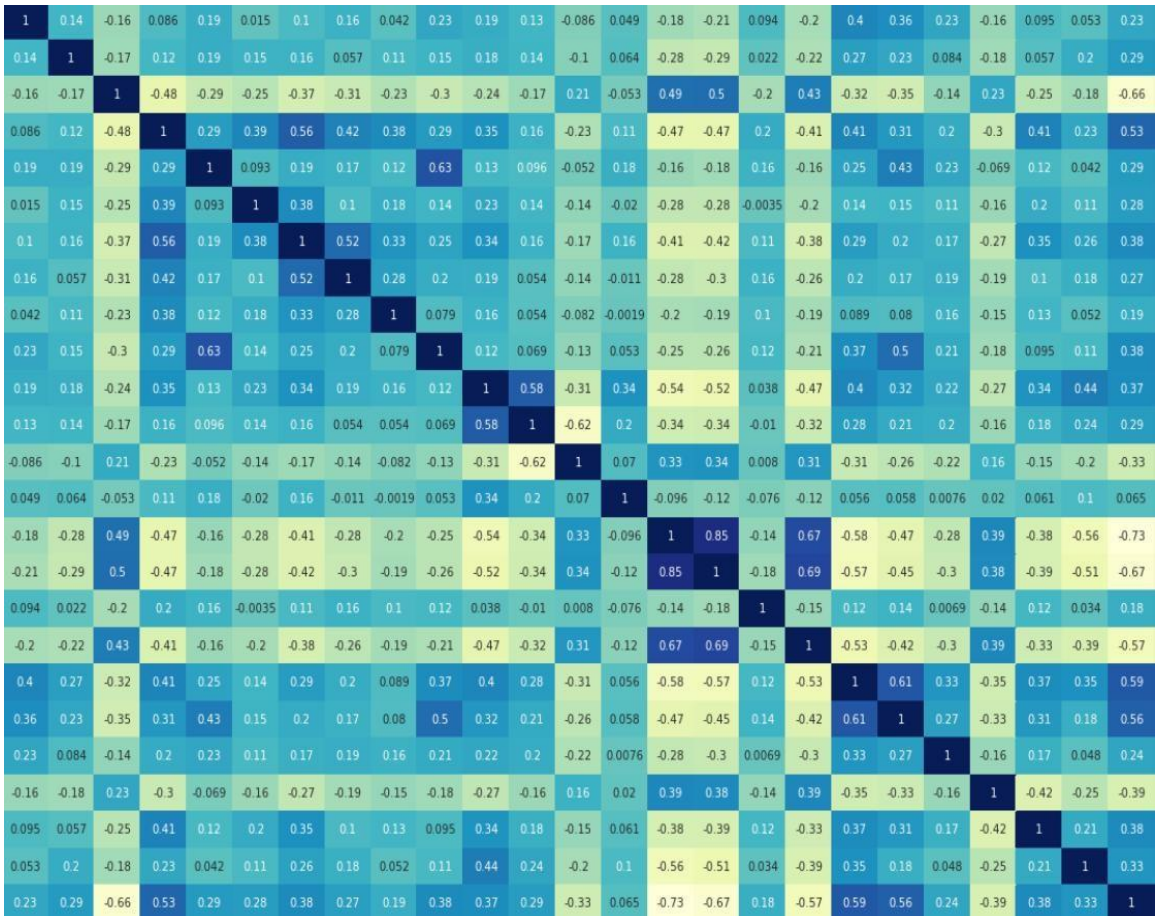


FIGURE 2. Heat map of chronic kidney

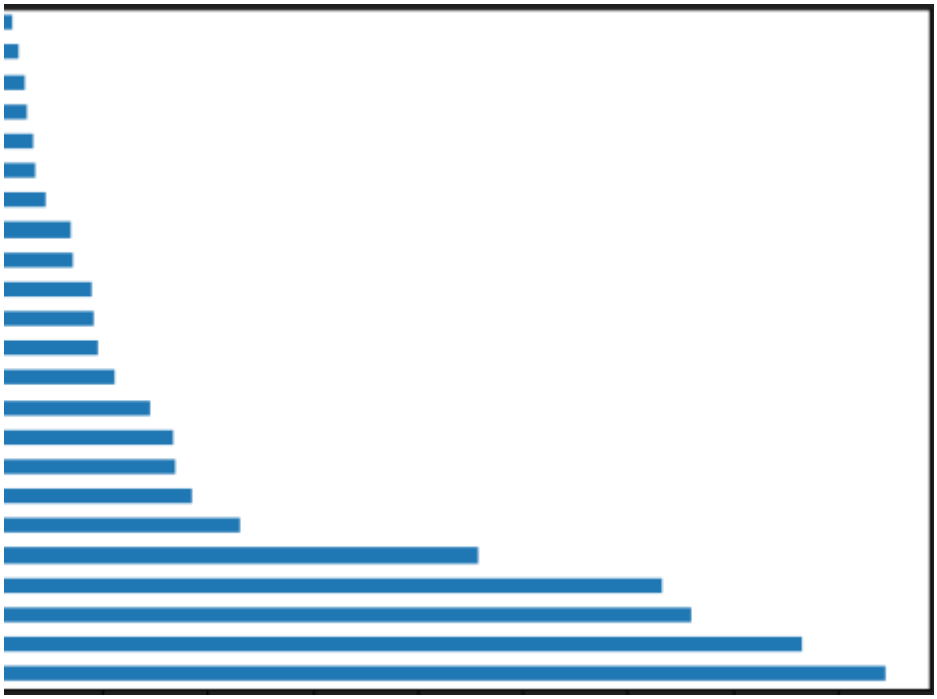


FIGURE 3. Feature Importance for CKD Prediction

TABLE 1. Chronic Kidney Disease

Evaluation Metric	Value
Accuracy	98.00%
Precision	90.50%
Recall	91.80%
F-1 Score	91.10%

7. CONCLUSION

The CKD prediction system demonstrates the potential of machine learning in improving early diagnosis and risk assessment for chronic kidney disease. By leveraging advanced classification models, robust data preprocessing, and an interactive user interface, the system provides accurate and interpretable predictions. The ability to process real-time patient data enhances its practical utility in healthcare settings. Despite challenges such as data imbalance and variability in medical records, the integration of techniques like SMOTE, feature selection, and explainable AI ensures the model remains reliable and effective. The system's batch processing capability further supports scalability for clinical applications. In conclusion, the CKD prediction system contributes to timely medical intervention, improving patient outcomes and aiding healthcare professionals in making informed decisions. Future enhancements could focus on integrating real-time electronic health records (EHRs), expanding the dataset for better generalization, and incorporating deep learning approaches for further accuracy improvements.

REFERENCES

[1]. Smith, J., Brown, A., & Patel, R. (2023). "Ensemble Learning Approaches for CKD Diagnosis." Computational Medicine, 18(4), 34-50.

[2]. Purushotham Reddy, M., Srinivasa Reddy, K., Lakshmi, L., Mallikarjuna Reddy, A. Effective technique based on intensity huge saturation and standard variation for image fusion of satellite images, International Journal of Engineering and Advanced Technology, 2019, 8(5), pp. 291–295

[3]. Srinivasa Reddy, K., Suneela, B., Inthiyaz, S., ... Kumar, G.N.S., Mallikarjuna Reddy, A. Texture filtration module under stabilization via random forest optimization methodology, International Journal of Advanced Trends in Computer Science and Engineering, 2019, 8(3), pp. 458–469

[4]. Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G. Generating cancelable fingerprint template using triangular structures, Journal of Computational and Theoretical Nanoscience, 2019, 16(5-6), pp. 1951–1955

[5]. Chandrasekhara Reddy, T., Pranathi, P., Mallikarjun Reddy, A., Vishnu Murthy, G., Kavati, I. Biometric template security using convex hulls features, Journal of Computational and Theoretical Nanoscience, 2019, 16(5-6), pp. 1947–1950

[6]. Mallikarjuna, A., Karuna Sree, B. Security towards flooding attacks in inter domain routing object using ad hoc network, International Journal of Engineering and Advanced Technology, 2019, 8(3), pp. 545–547

[7]. 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.

[8]. 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

[9]. 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

[10]. 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.

[11]. 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

[12]. 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.

[13]. 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.

Copyright@ REST Publisher

[14]. 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.
[15]. 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.
[16]. 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.
[17]. 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.
[18]. 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.
[19]. 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
[20]. 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.
[21]. 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.
[22]. 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
[23]. 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.
[24]. 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

Copyright@ REST Publisher