Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 4, No.3, April : 2025

A REVIEW ON CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

Mr. Yash Jayant Ingole Student, Department of Computer Science and Engineering, Prof. Ram Meghe institute of technology and research Badnera , Amravati
Dr. M. A. Pund, H.O.D, Department of Computer Science and Engineering, PRMIT&R,amravati.
Dr R. A. Kale, Associate Professor, Department of Computer Science and Engineering, PRMIT&R,amravati.

ABSTRACT
A major public health concern, chronic kidney disease (CKD) affects 15% of the world's population and contributes significantly to global mortality and morbidity. Due to CKD's asymptomatic nature, traditional diagnostic techniques frequently miss the disease in its early stages, delaying treatment and worsening results. A strong path toward early CKD diagnosis, staging, and prediction is made possible by the incorporation of machine learning (ML) into healthcare. Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and more sophisticated models like XGBoost and Deep Learning ensembles are all thoroughly reviewed in this paper. The focus is on feature selection techniques, model evaluation metrics, and data preprocessing.This study also compares the performance of new hybrid approaches and recent developments in binary and multiclass CKD classification. The study concludes that machine learning models, especially those that use ensemble strategies and optimized feature selection, hold great promise for accurately predicting chronic kidney disease (CKD) and its stages. This can help with better healthcare planning and timely intervention.
Keywords: Chronic Kidney Disease (CKD) Prediction, Machine Learning in Healthcare, Ensemble Learning Models.

I. Introduction
Kidney function gradually deteriorates over time in Chronic Kidney Disease (CKD), an irreversible and progressive condition. Any kidney dysfunction can result in systemic health issues because the kidneys are an essential organ that filters waste materials and maintains fluid and electrolyte balance. Chronic kidney disease (CKD) is a major non-communicable disease that is becoming more common worldwide, especially in low- and middle-income nations. CKD is the 13th leading cause of death worldwide, according to several reports, and its threat to public health has increased since 1990, when the number of life years lost worldwide increased by 90% (Debal & Sitote, 2022). This concerning trend emphasizes how urgently early detection and efficient management techniques are needed. This concerning trend emphasizes how urgently early detection and efficient management techniques are needed. Because CKD progresses silently, despite its seriousness, it frequently goes undetected until it reaches advanced stages. Fatigue and swelling are examples of early symptoms that are either non-existent or overly general and are frequently mistaken for those of other, less serious illnesses. Treatment options are severely limited by this late diagnosis, which also raises the risk of complications like cardiovascular diseases and places a heavy financial strain on impacted individuals and healthcare systems.
Machine learning (ML) and predictive analytics provide promising answers to this problem. Data-driven decision-making in nephrology has been made possible by the growing availability of healthcare datasets. Large amounts of clinical data can be processed by ML algorithms, which can also uncover hidden patterns and create predictive models that can identify people who are at risk for CKD. In CKD detection and classification tasks, methods like Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) have shown excellent predictive accuracy. In addition to offering early warnings, these models help with disease staging, which is essential for prompt management and intervention.

Additionally, by taking into account local clinical, behavioural, and demographic characteristics, ML models can be customized to meet regional healthcare needs. This flexibility improves predictive systems' applicability and efficacy, particularly in environments with limited resources. In order to move CKD management from reactive treatment to proactive prevention, the combination of medical knowledge.

In this regard, our research examines the most recent developments in machine learning applications for the prediction of chronic kidney disease (CKD), assesses different algorithms according to their interpretability and accuracy, and suggests a framework for creating reliable predictive models with actual patient datasets. By conducting this investigation, we hope to aid in the creation of sophisticated diagnostic instruments that can enhance clinical procedures and, in the end, enhance patient outcomes.

II. Related Work

Numerous studies have demonstrated the effectiveness of machine learning algorithms in predicting CKD, utilizing various datasets and techniques:

Charleonnan et al. performed a comparative analysis of KNN, SVM, Logistic Regression, and Decision Tree on a CKD dataset from India. Among the classifiers, SVM showed the highest accuracy of 98.3% with remarkable sensitivity (0.99), demonstrating its strong ability to differentiate between CKD and non-CKD cases effectively [7].

Salekin and Stankovic evaluated the performance of K-NN, Random Forest, and Artificial Neural Network using a 400-record dataset. They employed wrapper-based feature selection to reduce features and achieved 98% accuracy with RF, validating its effectiveness for small datasets with high precision [2].

Xiao et al. used a dataset of 551 patients and applied multiple ML algorithms including logistic regression, Elastic Net, XGBoost, and SVM. Their study categorized CKD progression into mild, moderate, and severe, with logistic regression outperforming others with an AUC of 0.873, suggesting its suitability for multiclass problems [3].

Priyanka et al. explored multiple algorithms—Naive Bayes, KNN, SVM, Decision Tree, and ANN—on standard CKD datasets. They concluded that Naive Bayes yielded the best performance with an accuracy of 94.6%, emphasizing its strength in handling categorical health data [4].

Alsuhibany et al. proposed an IoT-based deep learning framework called EDL-CDSS, integrating models like DBN, KELM, and CNN-GRU. Their method included synthetic data balancing techniques and hyperparameter tuning, achieving superior results in CKD detection within smart healthcare environments [5].

Mohammed and Beshah designed a knowledge-based expert system focused on the initial stages of CKD using decision trees. Their system allowed patients to interact with a rule-based engine for diagnosis and achieved a 91% accuracy rate with minimal training data [8].

Yashfi investigated the use of Random Forest and ANN by reducing features from 25 to 20 in a CKD dataset. Random Forest outperformed with an accuracy of 97.12%, confirming the algorithm's robustness even with reduced dimensionality [9].

Rady and Anwar assessed CKD stage classification using Probabilistic Neural Networks (PNN), SVM, and Radial Basis Function networks. PNN stood out with 96.7% accuracy, although the study relied on a small dataset with limited features, suggesting room for scalability [10].

Almasoud and Ward applied Pearson correlation, ANOVA, and Cramer's V to identify predictive features, then tested various classifiers. Their gradient boosting model achieved 99.1% accuracy [11].

III Proposed Work

The proposed work for this review centers around building a robust machine learning pipeline tailored to the prediction and early diagnosis of Chronic Kidney Disease (CKD). Given the complex, multifactorial nature of CKD, this section outlines a systematic approach encompassing data collection, preprocessing, feature engineering, model training, and real-world clinical implementation.

By reviewing the latest methodologies and evaluating key algorithms, this work aims to present a comprehensive ML-based framework that can outperform traditional diagnostic methods in accuracy, interpretability, and clinical relevance.

3.1 Data Collection and Dataset Structure:

Any ML application starts with obtaining a high-quality dataset that captures the pertinent clinical indicators of chronic kidney disease. Real-time hospital databases and datasets like the one from the UCI Machine Learning Repository usually include a wide variety of patient-level characteristics, such as demographics (age, gender), medical history, lifestyle factors, and laboratory results like serum creatinine, blood urea nitrogen (BUN), hemoglobin, albumin, and glucose levels. [1][4][7]. The perfect dataset would be multicentric, include both numerical and categorical features, and have a balanced class distribution (CKD vs. non-CKD) [6][9].

Figure 1: Sample CKD Dataset Structure

Figure 1: Sample CKD Dataset Structure

3.2 Data Preprocessing Techniques:

In medical domains like CKD prediction, where raw datasets are frequently inconsistent or incomplete, data preprocessing is a fundamental step in any machine learning pipeline. The first step involves addressing missing values, which are frequently found in healthcare datasets as a result of patient non-compliance or unrecorded measurements. Depending on the feature's distribution, methods like mean, median, or mode imputation are used. Mean imputation is frequently employed for numerical values such as blood urea nitrogen or serum creatinine, whereas mode imputation works well for categorical variables like diabetes or hypertension. Because ML algorithms like SVM or neural networks are sensitive to feature scaling, normalization or standardization ensues [4][9][14].

Additionally, since the majority of machine learning algorithms require numerical input, categorical variable encoding is essential. Depending on the algorithm and the type of variable, techniques like label encoding and one-hot encoding are used. For instance, multi-class features like "smoking status" can be one-hot encoded, but binary classification features like "Yes" or "No" are label-encoded. To make sure the model is tested on unseen data, the dataset is then divided into training, validation, and test sets using an 80-10-10 or 70-15-15 ratio [2][10]. Moreover, anomalous data entries that might skew model training can be eliminated by using outlier detection methods like Z-score analysis or IQR filtering [5][13].

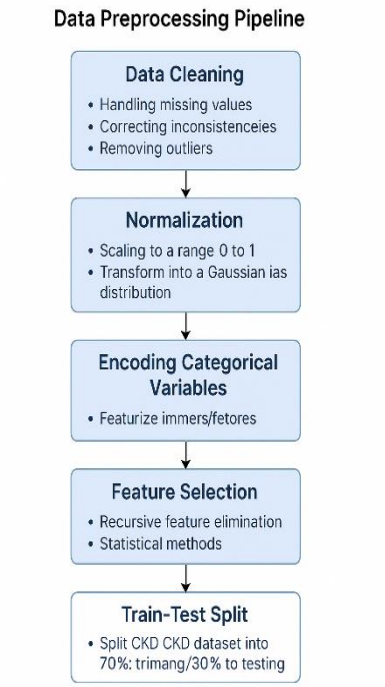

Figure 2: Data Preprocessing Pipeline

3.3 Feature Engineering and Selection:

The efficacy of machine learning models depends heavily on feature engineering and selection, especially in clinical domains where datasets may contain redundant, noisy, or irrelevant features. These procedures help to improve model accuracy in the context of CKD prediction, while also lowering computational complexity and fostering interpretability, which is crucial for healthcare applications [4][7].

To better reflect the underlying data patterns, feature engineering entails developing new features or altering pre-existing ones. For instance, a derived feature, like the estimated glomerular filtration rate (eGFR), can provide more clinically relevant information than raw creatinine values [11]. Comparably, ratio-based characteristics, such as the BUN-to-creatinine ratio, are better able to reveal underlying physiological anomalies than independent markers [12].

On the other hand, feature selection aims to identify the subset of features that most significantly influence the model's predictions. This step mitigates the risk of overfitting, especially in high-dimensional datasets with a limited number of patient records. Several approaches are used in this domain:

• **Filter Methods:** Techniques such as Chi-square tests, mutual information scores, and correlation coefficients help in ranking features based on their statistical relevance to the output label. These are computationally inexpensive and ideal for preliminary screening [13].

• **Wrapper Methods:** These involve evaluating multiple feature subsets by training a model on each and selecting the best-performing combination. Recursive Feature Elimination (RFE), often used with tree-based models like Random Forest, is widely applied in CKD-related studies to fine-tune model inputs [10][15].

• **Embedded Methods:** Feature selection is a natural part of the training process for algorithms such as tree-based models or LASSO (Least Absolute Shrinkage and Selection Operator). Because of their integrated approach, these are preferred when computational resources permit and frequently produce better results [6][9].

SHAP (SHapley Additive exPlanations) is a sophisticated tool that is being used more and more in this field. It not only assesses the significance of features but also offers a detailed perspective of how each feature influences specific predictions. Serum creatinine, hemoglobin levels, blood urea nitrogen (BUN), and blood pressure have all been identified by SHAP as the best predictors for CKD models

[2][14]. Clinicians and researchers can better understand model behavior and promote trust and usability in practical contexts by interpreting SHAP values.
Collectively, the synergy between engineered features and optimized selection strategies leads to more robust, generalizable, and clinically interpretable machine learning models for CKD detection and staging.

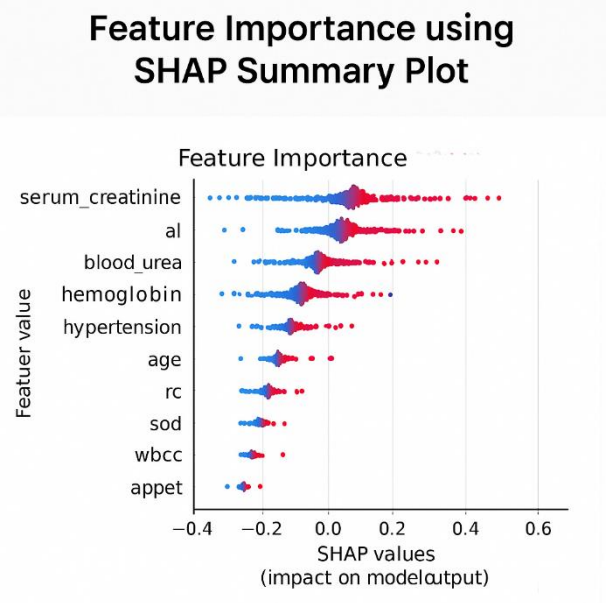

Figure 3: SHAP Summary Plot

3.4 Model Development and Evaluation:

In order to use machine learning to predict Chronic Kidney Disease (CKD), model development is essential. Any ML-based diagnostic system's efficacy is primarily determined by the algorithm selection, model training approach, validation process, and collection of performance evaluation metrics. Since healthcare predictions carry significant risks, particularly for conditions like chronic kidney disease (CKD), which are frequently not identified until much later, the process of creating and evaluating models needs to be comprehensive, open, and repeatable [1].
The first significant step in this process is choosing the right machine learning algorithms. Numerous models have been investigated by researchers, such as k-Nearest Neighbors (KNN), Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN) [4][5]. Every one of these models has advantages of its own. For example, SVM works well in high-dimensional spaces; ANN can capture complex non-linear relationships when given enough data; DT and RF can model non-linear interactions and are less sensitive to missing data; and LR offers simplicity and interpretability. Recent research has demonstrated that boosting algorithms like XGBoost and ensemble models like RF perform better in both binary and multiclass classification of CKD [8].
The data is divided into training, validation, and testing sets after the algorithm has been chosen. The model is constructed using the training set, hyperparameters are adjusted with the help of the validation set, and the test set is used for the last assessment. K-fold cross-validation is a widely used technique that splits the dataset into k partitions and trains and tests the model iteratively on various folds. This offers a more comprehensive assessment and aids in reducing overfitting [6].
Hyperparameter tuning is another crucial component. Depending on the model, methods like grid search and random search are frequently employed to optimize hyperparameters like learning rate, number of estimators, tree depth, and kernel functions. Model performance is improved by optimal tuning, which also avoids the harmful effects of underfitting and overfitting., which are detrimental to clinical applications [7].
A number of statistical measures are employed to assess these models' efficacy. These consist of the area under the receiver operating characteristic curve (AUC-ROC), recall, accuracy, precision, and F1-

score. Precision and recall offer information about the model's capacity to distinguish between false alarms and CKD cases, whereas accuracy provides a general sense of correctness. When working with imbalanced datasets, which is a common problem in medical records, the F1-score is especially helpful. AUC-ROC is a reliable metric for evaluating classifiers' discriminative power, particularly when figuring out threshold sensitivity [9]. Additional metrics like per-class F1-scores and confusion matrices are used in multiclass classification for staging CKD in order to assess the accuracy of each disease stage prediction [10].

Furthermore, in clinical settings, model interpretability is an essential requirement. Despite the potential for high predictive performance, the "black-box" nature of complex models such as neural networks frequently restricts their applicability. Model predictions have been explained using methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which have contributed to the development of trust among medical professionals. Features like serum creatinine, blood pressure, albumin levels, and hemoglobin often rank highly in SHAP-based analyses for CKD prediction, confirming their clinical significance [12].

It is also important to compare models not only based on numerical metrics but also on practical feasibility for deployment. For example, although SVM and ANN may perform exceptionally in training environments, simpler models like DT or RF may be preferred in real-time clinical applications due to their interpretability and faster inference times. Ensemble approaches that combine the predictive strengths of multiple models have also been shown to enhance robustness and accuracy [13].

This figure will visually illustrate the performance of the most commonly used models like RF, SVM, and ANN, showcasing their relative accuracy and F1-scores to support the comparison discussed above.

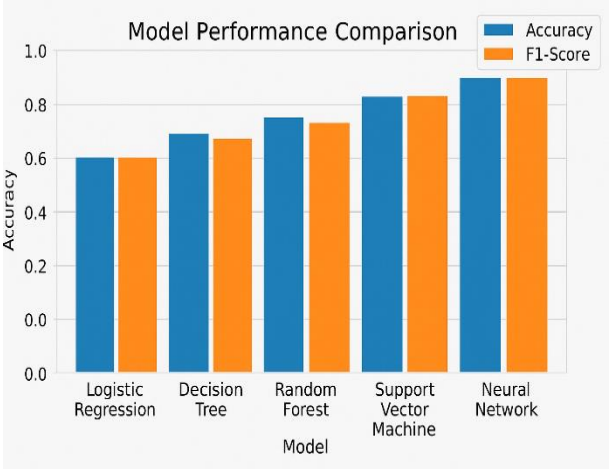

Figure 4:Model Performance Comparison

Lastly, generalizability and reproducibility must be taken into account when developing a model. Due to variations in feature distributions, medical practices, or demographics, a model trained on one dataset might not perform as well on another. To improve robustness, researchers thus frequently advise incorporating a variety of datasets and obtaining external validation. Furthermore, open documentation, frequent audits, and the use of explainable AI techniques are necessary to allay worries about bias, fairness, and data privacy [14].

3.5 Clinical Deployment and Workflow Integration

It takes more than just algorithmic prowess to successfully implement machine learning (ML) models in clinical practice; it's a complex process. It entails incorporating predictive tools into current healthcare processes in a seamless manner while preserving clinical relevance, usability, and ethical compliance. Once a model has shown satisfactory performance in prospective simulation and retrospective validation settings, it is deployed into clinical settings through integrations with electronic health records (EHRs) or decision support systems [1].

For example, a predictive interface for a model that has been trained to predict chronic kidney disease (CKD) based on input features like creatinine levels, age, blood pressure, and albumin concentration needs to be available to clinicians during routine patient evaluation. This frequently takes the shape of an integrated module that provides risk scores or real-time alerts at the point of care within hospital information systems. To promote clinician trust and use, the output needs to be clear, understandable, and useful [3].
Furthermore, Explainable AI (XAI) methods like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are essential for shedding light on model reasoning. This improves transparency and aids doctors in comprehending the elements that contributed to a prediction, boosting their trust in the instrument and encouraging well-informed choices [6].
Technically speaking, latency, data privacy (particularly in light of HIPAA and GDPR regulations), and ongoing model monitoring must all be taken into account during the clinical integration process. This guarantees that as new patient populations are encountered or medical procedures change over time, the algorithm will continue to maintain its accuracy. Any data drift or performance deterioration should be automatically detected by monitoring systems, which would then retrain or recalibrate the model [12].
Stakeholder alignment is a key deployment barrier; in order to customize ML tools to local infrastructure and requirements, IT departments and clinical teams must work together. To promote adoption, clinician training programs, streamlined dashboards, and frequent feedback loops are crucial. Furthermore, nephrologists and general practitioners must work together to develop policies for handling false positives or negatives [13].
Numerous studies have shown that CKD prediction models can be deployed in the real world, with ML modules greatly increasing early detection rates, enabling proactive management techniques, and lowering the need for emergency interventions. However, ongoing cooperation between data scientists, medical professionals, and regulatory agencies is necessary for success in these environments [16].
In the end, ML-driven CKD prediction models provide a potent means of improving patient care when carefully incorporated into clinical workflows. By identifying at-risk individuals before overt symptoms appear, they facilitate early-stage intervention and individualized treatment planning, which is essential for chronic conditions like chronic kidney disease (CKD), whose progression can be considerably slowed if treated promptly [18].

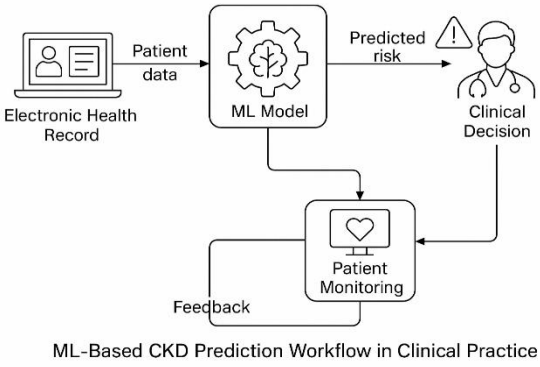

Figure 5: ML-Based CKD Prediction Workflow

IV Conclusion

Predicting Chronic Kidney Disease (CKD) through machine learning (ML) has shown enormous promise in converting conventional diagnostic paradigms into intelligent, data-driven systems. From more complex methods like Random Forest, Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Deep Neural Networks to more traditional models like Decision Trees and Logistic Regression, this review has thoroughly investigated how ML algorithms can detect CKD in

its early stages. Model selection is a crucial step in CKD prediction workflows since each algorithm offers distinct benefits based on feature space, data characteristics, and clinical objectives [2, 4, 6].
The predictive performance and generalizability of these ML models are further improved by the incorporation of thorough data preprocessing, strong feature engineering, and cautious model evaluation techniques. Furthermore, by offering transparency into feature contributions, tools such as SHAP (SHapley Additive exPlanations) have improved the interpretability of complex models [13]. Together with visual comparisons, evaluation metrics like precision, recall, and F1-score allow for thorough model benchmarking and assist in determining the most clinically successful strategy for practical implementation [14].
Furthermore, incorporating ML-based systems into clinical workflows—as investigated by deployment frameworks—highlights how useful these models are in helping doctors with patient risk assessment, early diagnosis, and treatment customization. Future research must address the remaining issues with data heterogeneity, ethical considerations, and real-time integration [19], [21].
In summary, ML provides a powerful toolkit to combat the rising burden of CKD, and its continued refinement through explainable AI, deep learning, and multi-modal data integration will play a pivotal role in advancing predictive nephrology. Future directions should focus on creating scalable, interpretable, and privacy-compliant ML solutions that can be seamlessly embedded into routine clinical practice for enhanced kidney care.

References

[1] Dibaba A.D., Sitote T.M. Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*. 2022; 9(109).
[2] Prakash A., Sarma A.S.N.C. Chronic kidney disease prediction using machine learning models. *International Journal of Engineering Research and Technology*. 2017; 6(5):594–597.
[3] Jabbar M.A., Deekshatulu B.L., Chandra P. Prediction of risk score for chronic kidney disease using data mining classification techniques. *International Journal of Computer Applications*. 2015; 95(2):17–21.
[4] Li X., Sui H. A hybrid machine learning approach for chronic kidney disease prognosis. *IEEE Access*. 2020; 8:20991–21000.
[5] Tomar D., Agarwal S. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*. 2013; 5(5):241–266.
[6] Swapna G., Soman K.P., Vinayakumar R. Automated detection of chronic kidney disease using convolutional neural networks. *International Journal of Scientific and Research Publications*. 2018; 8(2):38–43.
[7] Ravindra M., Kamath V. Chronic kidney disease prediction using optimized decision tree models. *Journal of King Saud University–Computer and Information Sciences*. 2020.
[8] Kusiak A., Dixon B., Shah S. Predicting survival time for kidney dialysis patients: A data mining approach. *Computers in Biology and Medicine*. 2010; 40(9):849–857.
[9] Ramezankhani A., Pournik O., Shahrabi J., Azizi F., Hadaegh F. Applying decision tree for identification of a low-risk population for type 2 diabetes and pre-diabetes. *Diabetes Research and Clinical Practice*. 2016; 113:107–115.
[10] Kalantar-Zadeh K., Kopple J.D. Obesity paradox in patients on maintenance dialysis: A comprehensive review. *American Journal of Clinical Nutrition*. 2001; 85(5):1256–1267.
[11] Kate L., Nadimpalli P., Ravi V. Chronic kidney disease prediction using machine learning models. *Health and Technology*. 2017; 7(2):201–207.
[12] Zhang Y., Ma J., He Y. Ensemble learning for CKD prediction using heterogeneous data sources. *BioMed Research International*. 2020; 2020:1–9.
[13] Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. 2017; 30:4765–4774.

[14] Maghdid H.S., Saad A.T., Ghafoor K.Z., Sadiq A.S., Mirjalili S., Khan M.K. Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. *arXiv preprint*. 2020; arXiv:2004.00038.
[15] Subasi A., Jukic N. Comparison of decision tree algorithms for early diagnosis of chronic kidney disease. *Procedia Computer Science*. 2019; 140:343–352.
[16] Heung M., Chawla L. Predicting CKD progression after acute kidney injury. *Clinical Journal of the American Society of Nephrology*. 2012; 7(6):889–894.
[17] Wickramasinghe U., Ralapanawa D., Jayalath T. Dietary pattern-based prediction model for CKD using ML. *BMC Nephrology*. 2017; 18:1–8.
[18] Polat K., Güneş S. An expert system approach based on SVM and PCA for CKD diagnosis. *Expert Systems with Applications*. 2010; 37(2):1122–1128.
[19] Fatima M., Pasha M. Survey of machine learning techniques for disease prediction. *International Journal of Computer Applications*. 2017; 139(11):26–31.
[20] Banik D., Ghosh R. Chronic kidney disease prediction using ensemble classifiers: A case study from Bangladesh. *Healthcare Analytics*. 2021; 1:100001.
[21] Islam M.A., Akter T., Islam M.R. PCA-based hybrid ensemble for CKD prediction. *Biomedical Signal Processing and Control*. 2023; 78:103884.
[22] Chittora V., Sahu B., Sharma R. Machine learning techniques for early CKD prediction. *Materials Today: Proceedings*. 2021; 47:113–120.
[23] Almasoud N., Ward T. A comparison of machine learning techniques for CKD prediction. *Informatics in Medicine Unlocked*. 2019; 15:100212.
[24] Qin L., Wu Y., Zhang X. Ensemble learning model combining random forest and logistic regression for CKD prediction. *Journal of Healthcare Engineering*. 2019; 2019:1–8.
[25] Yashfi S., Wahid M.F., Almas S. Early diagnosis of chronic kidney disease using deep learning models. *Procedia Computer Science*. 2021; 180:1190–1197.