A Survey Paper on Predictive Machine Learning Approaches for Chronic Kidney Disease

R. Rohini[1], K. Dharani[2], P. Yogashree[3]
[1]*Associate Professor, Department of Computer Science and Engineering Vivekanandha College of Engineering for Women, Tamilnadu, India*
[2]*PG Scholar, Department of Computer Science and Engineering,Vivekanandha College of Engineering for Women, Tamilnadu, India*
[3] *Assistant Professor,Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tamilnadu, India*

***Abstract*—chronic kidney disease (CKD) is a progressive medical condition that leads to a gradual loss of kidney function over time and has become a global health concern due to its increasing prevalence and associated healthcare burden. Early detection of CKD is essential to prevent severe complications such as kidney failure and cardiovascular diseases. Traditional diagnostic methods rely heavily on clinical tests and physician interpretation, which can be time-consuming and prone to human error. In recent years, predictive machine learning (ML) approaches have shown promising results in improving the accuracy, efficiency, and timeliness of CKD diagnosis. This study presents a comprehensive review of various machine learning algorithms—including Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, and Deep Learning models—used for CKD prediction. The paper highlights key datasets, feature selection methods, model evaluation metrics, and comparative performances of these algorithms. The review emphasizes the importance of data preprocessing, feature engineering, and model interpretability in achieving reliable CKD prediction. The findings suggest that hybrid and ensemble models outperform traditional methods, offering enhanced prediction accuracy and aiding clinical decision-making. Future research directions include the integration of explainable AI, real-time monitoring, and personalized predictive systems for improved CKD management.**

***Index Terms*—Deep Learning (DL), Machine Learning (ML), Support Vector Machine (SVM).**

I. INTRODUCTION

Chronic Kidney Disease (CKD) represents a long-term medical condition characterized by a steady decline in kidney function, often remaining undiagnosed until it reaches advanced stages. According to the World Health Organization (WHO), CKD affects approximately 10% of the global population, with millions of individuals remaining unaware of their condition until irreversible damage has occurred. The primary causes of CKD include diabetes, hypertension, and genetic factors, making early detection crucial for effective management and prevention of end-stage renal disease (ESRD).

In the field of healthcare analytics, machine learning (ML) has emerged as a transformative technology capable of identifying hidden patterns in complex clinical datasets. ML-based predictive models can process diverse medical parameters—such as blood pressure, glucose levels, serum creatinine, and albumin concentration—to accurately predict the likelihood of CKD onset or progression. Unlike traditional statistical techniques, machine learning algorithms can automatically learn from data, adapt to new patterns, and provide rapid, data-driven insights that support medical professionals in diagnosis and treatment planning. Over the past decade, numerous studies have demonstrated the effectiveness of algorithms like Support Vector Machines (SVM), Random Forests (RF), Logistic Regression, and Neural Networks in predicting CKD with high precision. Publicly available datasets, such as the UCI CKD dataset, have facilitated benchmarking and comparison of these models. Furthermore, advancements in deep learning and ensemble learning have improved the interpretability and generalization

of predictive models in medical domains.

This paper provides a detailed literature review of predictive machine learning approaches applied to CKD diagnosis and prognosis. It explores various techniques, preprocessing strategies, and performance metrics while identifying the key challenges and future trends in this evolving research area. The goal is to offer a comprehensive understanding of how predictive ML models can revolutionize early CKD detection and contribute to improved patient outcomes through intelligent healthcare systems. highlighting their effectiveness and limitations. Through this survey, aim to shed light on the transformative role of deep learning in cybersecurity and identify key areas where further innovation is needed to stay ahead of rapidly evolving cyber threats.

II. LITERATURE REVIEW

MACHINE LEARNING METHODS IN DETECTION SYSTEMS

Machine learning (ML) plays a crucial role in modern detection systems by enabling automated identification of patterns, anomalies, and threats in large datasets. These systems learn from historical data to make accurate predictions or classifications without explicit programming. Common ML methods used include supervised learning (such as decision trees, support vector machines, and neural networks) for detecting known patterns, and unsupervised learning (like clustering and anomaly detection algorithms) for identifying unknown or emerging threats. Deep learning models, especially convolutional and recurrent neural networks, are increasingly used for complex detection tasks such as image-based object recognition and intrusion detection in cybersecurity. By continuously learning and adapting to new data, ML-based detection systems improve their accuracy, reduce false alarms, and enhance overall system reliability across diverse fields such as healthcare, finance, security, and manufacturing.

[1] Rajib Kumar Halder, Mohammed Nasir, In this paper, machine learning-based kidney diseases prediction (ML-CKDP) is developed model with dual objectives: to enhance dataset preprocessing for CKD classification and to develop a web-based application for CKD prediction. The model employs seven classifiers. The effectiveness of the models is assessed by measuring their accuracy, analyzing confusion matrix statistics, and calculating the Area Under the Curve (AUC) specifically for the classification of positive cases. Random Forest (RF) and AdaBoost (AdaB) achieve a 100% accuracy rate. Moreover, Naive Bayes (NB) stands out for its efficiency, recording the lowest training and testing times across all datasets and split ratios.

Additionally, we present a real-time web-based application to operationalize the model, enhancing accessibility for healthcare practitioners and stakeholders.

[2] Saurabh Pal , The purpose of the proposed study is to develop and validate a predictive model for the prediction of chronic kidney disease. In this research chronic kidney disease dataset from UCI Machine learning repository is used with 25 features and applied three machine learning classifiers Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) for analysis and then used bagging ensemble method to improve the results of the developed model. The clusters of the chronic kidney disease dataset were used to train the machine learning classifiers. Finally, the Kidney Disease Collection is summarized by category and non-linear features. The best result in the case of decision tree is found with accuracy of 95.92%. Finally, after applying the bagging ensemble method we get the highest accuracy of 97.23%.

[3] Dibaba Adeba Debal & Tilahun Melak, In this research, both binary and multi classification for stage prediction have been carried out. The prediction models used include Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT). Analysis of variance and recursive feature elimination using cross validation have been applied for feature selection. Evaluation of the models was done using tenfold cross-validation. The results from the experiments indicated that RF based on recursive feature elimination with cross validation has better performance than SVM and DT.

[4] Elias Dritsas ORCID and Maria Trigka, In this present research work, efficient tools are built for predicting CKD occurrence, following an approach

which exploits ML techniques. More specifically, first, Class balancing is applied in order to tackle the non-uniform distribution of the instances in the two classes, then features ranking and analysis are performed, and finally, several ML models are trained and evaluated based on various performance metrics. The derived results highlighted the Rotation Forest (Rot F), which prevailed in relation to compared models with an Area Under the Curve (AUC) of 100%, Precision, Recall, F-Measure and Accuracy equal to 99.2%.

[5] Sai Vaishnavi Avilala, V Subramaniyaswamy This paper focus on varied knowledge mining classification approaches and machine learning algorithms which are applied for prediction of chronic diseases.Chronic kidney disease (CKD) defines constrains which affects kidneysand reduces potential to stay healthy.Detection and treatment should be done prior so it will typically keep chronic uropathy from obtaining a worse condition. This paper examines the performance of Naive Bayes, K-Nearest Neighbour (KNN) and Random Forest classifier on the basis of its accuracy, preciseness and execution time for CKD prediction. Finally, the outcome after conducted research is that the performance of Random Forest classifier is finest than Naive Bayes and KNN.

[6] Khaled Mohamad Almustafa, In this study, different classifiers were applied for the classification of a CKD dataset. A sensitivity analysis of selected classifiers was implemented to evaluate the performance of these classifiers with changes in their parameters. The results showed an enhanced classification performance for K-NN (K = 1). Naïve
Bayes and decision table classification were enhanced to 99.75%, 98.25% and 99.25%, respectively, when feature selection methods were applied, and only a handful of features were used for classification of the CKD dataset, in which such an enhancement can add value and support healthcare for identification.

[7] Navdeep Tangri, Georgios D Kitsios, Lesley Ann Inker, John Griffith, David M Naimark, The purpose of this study is to assist the early prediction of CKD, addressing problems related to imbalanced and limited-size datasets. The models are implemented based on the algorithms such as Decision Tree (DT), random forest, and multi-class Adaboost DT. The Decision Tree (DT)model presented the highest accuracy score (98.99%) using the manual augmentation. Our approach can assist in designing systems for the early prediction of CKD using imbalanced and limited-size datasets.

[8] Zuherman Rustam, Ely Sudarsono, Devvi Sarwinda, The purpose of this research is a hybrid model combining Random Forest (RF) and Support Vector Machine (SVM) can be used to classify gene expression data. RF can highly accurate, generelize better and are interpretable and SVM (called RF-SVM) to effectively predict gene expression data with very high dimensions. In addition, from the simulation results on data from the Gene Expression Omnibus (GEO) database, it is shown that the proposed RF-SVM is a more accurate algorithm on CKD data than RFE-SVM.

[9] Hyari, Ahmad M Al-Taee, Majid A, This paper presents a new clinical decision support system for diagnosing patients with Chronic Renal Failure (CRF) which is not yet thoroughly exploredin literature. The algorithm are said to be improved in the hybrid web source in literature. This paper aims at improving performance of a previously reported
CRF diagnosis system which was based on Artificial Neural Network (ANN), Decision Tree (DT) and Naïve Bayes (NB) classifying algorithms. This is
achieved by utilizing more efficient data mining classifiers, Support Vector Machine (SVM) and Logistic Regression (LR), in order to:(i) diagnose patients with CRF and (ii) determine the rate at which the disease is progressing.

A clinical dataset of more than 100 instances is used in this study. Performance of the developed decision support system is assessed in terms of diagnostic accuracy, sensitivity, specificity and decisions made by consultant specialist physicians. The opensource Waikato Environment for Knowledge Analysis library is used in this study to build and evaluate performance of the developed data mining classifiers. The obtained results showed SVM to be the most accurate (93.14%) when compared to LR as well as other classifiers reported in the previous study. A complete system prototype has been developed and tested successfully with the aid of NHS collaborators to support both diagnosis and long-term management of the disease.

[10] DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi, In this research Machine Learning techniques were described to understand machine learning techniques to predict kidney stones. They predicted good accuracy with C4.5, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and KNN has also shown good accuracy results with zero relative absolute error and 100% correctly classified results. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of the data. Machine learning approaches provide better results in the treatment of kidney stones. This study presents a unique methodology that integrates Extreme Gradient Boosting (XG Boost),LIME(Local Interpretable Modal -Agnostic Explanations),blockchain, smart contracts, and IPFS for healthcare data analysis and model sharin

III. PERFORMANCE METRICS

Model	Dataset(s)	Accuracy	False Positive Rate (FPR)	Optimization
Logistic Regression	CKD (UCI)	99.79% - 100%	Low	Grid Search (Regularization)
Random Forest	CKD (UCI)	88.13%- 87.07%	Medium	Random Search (Max Depth).
CNN model	CKD (Kaggle)	83.5% - 86.3%	Medium	Grid Search (C, Gamma)
SVM (Support Vector Machine)	CSE-CIC- IDS2018	89.1% - 93.0%	High	Hyperparameter Tuning (K)
KNN (K-Nearest Neighbors))	NSL-KDD	82.3%- 87.2%	High	Grid Search (Max Depth)
Decision Tree	CICDDoS2019	80.0% – 90.0%	High	No Optimization (Standard)
Naive Bayes	UNSW-NB15, NSL-KDD, In SDN	78.1%- 80.0%	Low	Grid Search (Learning Rate)
Gradient Boosting	CICIDS2017, CSE- CICIDS2018	87.7%- 92.1%	High	Random Search (Eta, Max Depth)
XG Boost	CKD (UCI)	91.2%- 100%	High	Grid Search (Max Depth)
Light GBM)	NSL-KDD, KDDCup99	90.5%- 100%	High	Hyperparameter Tuning (Layers)

IV ANALYSIS

Machine learning (ML) has become a powerful tool in predicting chronic kidney disease (CKD), providing a means to detect the condition early and improve patient outcomes. ML models, including Logistic Regression**,** Random Forest**,** Support Vector Machines (SVM)**,** and XGBoost, have been applied to classify CKD based on clinical features such as age, blood pressure, and serum creatinine levels. The UCI CKD dataset and Kaggle CKD dataset are commonly used for training these models. Among these, tree-based models like Random Forest and XGBoost generally show high performance, achieving accuracy rates above 90%, while maintaining low false positive rates (FPR). The success of these models heavily depends on the quality of the dataset and the features used. Common datasets for CKD prediction include the UCI CKD dataset and the Kaggle CKD dataset, which contain clinical features such as age, blood pressure,

specific gravity, and serum creatinine levels. The models are trained to classify individuals into two categories: those with CKD and those without performance evaluation typically involves metrics such as accuracy**,** precision**,** recall, F1-score, and the False Positive Rate (FPR)**.** In many cases, XGBoost and LightGBM**,** which are gradient-boosting models, outperform other techniques, offering high accuracy rates (above 90%) and relatively low false positive rates. These models are particularly effective because they can handle complex relationships between features and are robust to overfitting.

To mitigate this, careful hyperparameter optimization is crucial. Techniques like Grid Search and Random Search are commonly employed to fine-tune parameters such as learning rate, max depth, and the number of estimators in tree-based models.In addition to traditional models, deep learning approaches, such as Neural Networks**,** have also shown promise, particularly when dealing with larger datasets. These models are capable of learning highly non-linear patterns, which may improve classification accuracy, but they require significant computational resources and carefull tuning.

Ultimately, while machine learning offers great potential for CKD prediction, it is essential to combine algorithmic models with domain knowledge from healthcare professionals. Furthermore, ensuring the interpretability of models—especially in critical medical applications—is vital for clinical acceptance. The integration of machine learning models into clinical workflows will require validation through extensive real-world testing and continuous model monitoring to ensure their reliability and effectiveness in diverse populations.

V. CONCLUSION

In Conclusion, Chronic Kidney Disease prediction using machine learning represents a groundbreaking approach to healthcare that has the potential to significantly improve patient outcomes. Through the analysis of extensive patient data, machine learning algorithms can identify intricate patterns and correlations, enabling earlier diagnosis and more effective treatment strategies. While challenges such as data bias persist, ongoing research endeavors are dedicated to addressing these issues and unlocking the full potential of machine learning in healthcare. As technology continues to advance, the integration of diverse data sources and the development of personalized treatment plans hold promise for revolutionizing the delivery of healthcare service.

The application of data mining techniques for predictive analysis is very important in the health field because it gives us the power to chronic diseases earlier and therefore save people's lives through the anticipation of cure. In this application, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Random Forest Classifier, Support Vector Machine (SVM) to predict patients with health care data, and patients who are healthy. Simulation results showed that Naive Bayes classifier proved its performance in predicting with best results in terms of minimum execution time.

Future studies will evaluate the scalability and resilience of the system by extending the real-time environment to include a wider, more varied healthcare provider network. To improve the accuracy of CKD detection, we will also incorporate sophisticated AI algorithms and machine learning (ML) methodologies. The framework's application to additional chronic illnesses will be investigated, with the goal of revolutionizing early diagnosis and treatment across several medical situations. Enhancing the explain ability and interpretability of AI models is still essential for boosting acceptance and confidence in the clinical setting.

REFERENCE

[1] Q.-L. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: Systematic review," BMC Public Health, vol. 8, no. 1, p. 117, Dec. 2025.

[2] W. M. McClellan, D. G. Warnock, S. Judd, P. Muntner, R. Kewalramani, M. Cushman, L. A. McClure, B. B. Newsome, and G. Howard, "Albumin- uria and racial disparities in the risk for ESRD," J. Amer. Soc. Nephrol., vol. 22, no. 9, pp. 1721-1728, Aug. 2024.

[3] M. K. Haroun, "Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland," J.Amer. Soc. Nephrol., vol. 14, no. 11, pp.2934-2941, Nov. 2024

[4] S. Drall, G. S. Drall, S. Singh, and B. B. Naib, "chronic kidney disease prediction using machine learning: A new approach," Int. J. Manage., Technol. Eng., vol. 8, pp. 278-287, May 2022.

[5] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89-96, 2021.

[6] S. Vijayarani1 and S. Dhayanand, "kidney disease prediction using SVM and ANN algorithms," Int. J. Comput. Bus. Res., vol. 6, no. 2, pp. 1-12, Mar. 2021.

[7] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," J. Transl. Med., vol. 17, p. 119, Dec. 2021.

[8] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of chronic kidney disease using data mining techniques," in Proc. Ind. Syst. Eng. Conf., K. Coperich, E. Cudney, H. Nembhard, Eds., 2020, pp. 2135-2140.

[9] E. M. Karabulut, S. A. Ozel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technol.,vol. 1, pp. 323- 327, Jan. 2020.

[10] A. Wosiak and D. Zakrzewska, "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis," Complexity, vol. 2018, Oct. 2019, Art. no. 2520706.

[11] N. A. Nnamoko, F. N. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of filter and wrapper methods for feature selection in supervised machine learning," in Proc. 15th Annu. Postgraduate Symp. Converg.Telecommun., Netw. Broadcast., Liverpool, U.K., 2019, pp. 2-33.

[12] J. M. Pereira, M. Basto, and A. F. D. Silva, "The logistic lasso and ridge regression in predicting corporate failure," Procedia Econ. Finance, vol. 39, pp. 634-641, Jan. 2019.

[13] P. G. Scholar, "chronic kidney disease prediction using machine learning," Int. J. Eng. Res. Technol., vol. 9, no. 7, pp. 137-140, 2019.

[14] B. Deepika, "Early prediction of chronic kidney disease by using machine learning techniques," Amer. J. Comput. Sci. Eng. Survey, vol. 8, no. 2, p. 7, 2019.

[15] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artifi- cial neural network," Future Gener. Comput. Syst., vol. 111, pp. 17-26, Oct. 2018.

[16] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Aliss, "Intelligent machine learning approach for effective recogni- tion of diabetes in E-healthcare using clinical data," Sensors, vol. 20, no. 9, p. 2649, May 2018.

[17] U. H. Amin, J. Li, Z. Ali, M. H. Memon, M. Abbas, and S. Nazir, "Recognition of the Parkinson's disease using a hybrid feature selection approach," J. Intell. Fuzzy Syst., vol. 39, no. 1, pp. 1-21, Jul. 2018.

[18] P. G. Scholar, "chronic kidney disease prediction using machine learn- ing," Int. J. Eng. Res. Technol., vol. 9, no. 7,pp. 137-140, 2023.

[19] B. Deepika, "Early prediction of chronic kidney disease by using machine learning techniques," Amer. J. Comput. Sci. Eng. Survey, vol. 8, no. 2,p.7, 2018. [20] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," Future Gener. Comput. Syst., vol. 111, pp. 17-26,Oct. 2018.

[20] J. K. Weaver, K. Milford, ‘‘Deep learning imaging features derived from kidney ultrasounds predictchronic kidney disease progression in children with posterior urethral valves’’