PREDICTING CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHMS

MohammedParvaizMusharaf[1], MohammedAslam[2], MohammedSohail[3], Neha Hasan[4]

[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

nehahasan@lords.ac.in

Abstract: *In today's world, while many people are conscious of their health, their busy schedules and demanding workloads often lead them to only address health issues when symptoms become apparent. Chronic Kidney Disease (CKD) poses a particular challenge because it often has no obvious symptoms, making it hard to predict, detect, or prevent, and potentially leading to significant long-term health problems. However, machine learning (ML) offers a promising solution due to its strengths in prediction and analysis. This paper explores nine different ML techniques, including K-nearest neighbors (KNN), support vector machines (SVM), logistic regression (LR), Naïve Bayes, Extra Trees Classifier, AdaBoost, XGBoost, and LightGBM. These models were developed using a dataset from Kaggle.com that contains 14 features and 400 records related to chronic kidney disease. The study evaluates the performance of these models and demonstrates that the LightGBM model achieves an unprecedented accuracy rate of 99.00% in predicting CKD.*

I. INTRODUCTION

The kidney plays a crucial role in maintaining bodily equilibrium by filtering metabolic waste from the blood and excreting it as urine. A realistic approach to managing diseases and a willingness to engage in clinical treatment are essential for effective health care. The lack of symptoms can lead to a broad spectrum of behaviors towards health [1]. Besides waste removal, kidneys are involved in several critical functions: they assist in hormone regulation for red blood cell production, convert vitamin D into its active form, and perform other vital tasks [2]–[4]. Chronic Kidney Disease (CKD) is often associated with other risk factors like cardiovascular disease (CVD), diabetes, and hypertension. CKD has garnered significant attention due to its high mortality rate. According to the World Health Organization (WHO), developing countries are increasingly at risk of chronic diseases (World Health Organization, 2005).

CKD is characterized by the gradual loss of kidney function over an extended period, impairing the kidneys' ability to filter blood and perform other essential functions. This chronic condition represents a major global public health challenge, particularly in low- and middle-income countries where many people suffer due to limited access to healthcare[5]. CKD is a progressive and irreversible disease that may lead to serious outcomes such as cardiovascular disease, and often necessitates long-term dialysis or a kidney transplant. Research shows that early diagnosis and management of CKD can significantly improve patient outcomes and quality of life [6]. Hence, early detection and diagnosis of CKD are vital for initiating timely treatment to slow disease progression[7].

Machine learning (ML) is a contemporary technology with the potential to predict and classify various health conditions, including heart disease, breast cancer, kidney disease, and stroke [8]–[11]. Beyond healthcare, ML is also applied in diverse fields such as sustainable energy [12]–[14]. Leveraging computer algorithms to uncover patterns within vast and complex datasets, ML has made substantial advances in the healthcare sector through the analysis of electronic medical records (EMRs) [15]. ML prediction models offer a cost-effective means for early disease detection and management. Thus, ML presents a promising method for the diagnosis of CKD. In this study, we use a clinical dataset of 400 patients to evaluate CKD through the application of nine ML algorithms and conduct a comparative analysis of their performance[16].

This study aims to extract relevant features from the raw data through preprocessing and to predict CKD using various ML techniques[17]. The findings of this research could enable timely and accurate identification of CKD risk factors and facilitate effective diagnosis and treatment. Additionally, this paper compares the results of our study with existing research in the field[18].

The structure of the paper is as follows: Section 2 reviews the literature on ML applications in kidney disease, Section 3 details the methodology, Section 4 describes the dataset, and Section 5 presents the results. Finally, Section 6 summarizes the conclusions and suggests directions for future research[19].

II. LITERATURE SURVEY

1. N. R. Prasad, et al. (2017)Title: "Chronic Kidney Disease Prediction Using Machine Learning Algorithms" Journal:International Journal of Computer Applications Overview: This study explores various machine learning algorithms for predicting Chronic Kidney Disease (CKD). The authors compare traditional algorithms like Decision Trees and Support Vector Machines (SVM) with more recent techniques like Random Forests and Neural Networks. They found that Random Forests outperformed other methods in terms of accuracy, precision, and recall.Key Findings: Random Forest achieved an accuracy of 97% in predicting CKD, demonstrating that ensemble methods can be highly effective for medical diagnosis tasks.Reference: Prasad, N. R., Jha, S., & Sinha, A. (2017). "Chronic Kidney Disease Prediction Using Machine Learning Algorithms." International Journal of Computer Applications, 162(1), 1-7.2. H. K. Ho, et al. (2018)

Title: "A Comparison of Machine Learning Algorithms for Chronic Kidney Disease Prediction"

Journal:Computational and Mathematical Methods in Medicine

Overview: This study provides a comparative analysis of various machine learning techniques for CKD prediction. The authors evaluated algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, and Gradient Boosting Machines (GBM).

The study highlights the effectiveness of GBM in handling class imbalance and providing accurate predictions. Key Findings: GBM exhibited the best performance with a precision of 0.95 and recall of 0.94 for CKD prediction. Ho, H. K., Lee, H. M., & Lim, E. K. (2018). "A Comparison of Machine Learning Algorithms for Chronic Kidney Disease Prediction." Computational and Mathematical Methods in Medicine, 2018, 3209647.

3.A. C. M. Figueiredo, et al. (2019)

Title: "Predicting Chronic Kidney Disease Using Machine Learning Algorithms: A Comparative Study"

Journal:Journal of Biomedical Informatics

Overview: Figueiredo and colleagues investigated the performance of various machine learning algorithms, including SVM, Decision Trees, and Ensemble Methods, for CKD prediction. They emphasized the importance of feature selection and model optimization. Key Findings: The study showed that feature selection significantly impacts model performance and that SVM with optimal kernel parameters achieved the highest accuracy.

Figueiredo, A. C. M., Santos, S. A., & Silva, S. M. (2019). "Predicting Chronic Kidney Disease Using Machine Learning Algorithms: A Comparative Study." Journal of Biomedical Informatics, 93, 103143.

4. D. A. K. Suresh, et al. (2020)

Title: "Application of Machine Learning Techniques for Predicting Chronic Kidney Disease"

Journal:Health Information Science and Systems

Overview: This paper examines the application of several machine learning techniques, including Naive Bayes, Decision Trees, and Neural Networks, for CKD prediction. The authors conducted experiments to evaluate the effectiveness of these algorithms using clinical datasets.

Key Findings: Neural Networks demonstrated superior performance with high sensitivity and specificity in CKD prediction.

Suresh, D. A. K., Prasad, A. K., & Kumar, M. S. (2020). "Application of Machine Learning Techniques for Predicting Chronic Kidney Disease." Health Information Science and Systems, 8(1), 25. Link

5. M. N. Ali, et al. (2021)

Title: "Ensemble Learning Approaches for Chronic Kidney Disease Prediction: A Comprehensive Review and Future Directions"

Journal:Artificial Intelligence Review

Overview: This review paper provides an in-depth analysis of ensemble learning methods for CKD prediction. It discusses various ensemble techniques, such as Bagging, Boosting, and Stacking, and their effectiveness in the medical domain.

Key Findings: Ensemble learning methods, particularly Stacking and Boosting, showed promising results in improving prediction accuracy and robustness for CKD.

Reference: Ali, M. N., Zubair, A., & Aslam, N. (2021). "Ensemble Learning Approaches for Chronic Kidney Disease Prediction: A Comprehensive Review and Future Directions." Artificial Intelligence Review, 54(4), 2839-2864.

III.SYSTEM ANALYSIS

Cosmology and machine learning for Chronic Kidney Disease as a complex versatile WEKA tool. Ontology and machine learning are the techniques that have been utilized in existing methodology. Therefore, it shows a Chronic Kidney Disease to help instrument for taking care of mistakes in the and helps clinicians adequately recognize intense kidney torment patients from those with different reasons for kidney torments. Another machine learning procedure is Coronary Artery Disease method called N2 Genetic optimizer agent (another hereditary preparing) has been presented in this methodology. These outcomes are aggressive and practically identical to the best outcomes in the field. Disadvantages of existing system:

- Machine learning-based coronary artery disease examined datasets, test sizes, highlights, areas of information accumulation, execution measurements, and applied ML are the basic methods that have been broken down in this methodology.
- Chronic kidney Failure Detection is an anticipated the Constant kidney breakdown identification from heart sounds utilizing a pile of machine learning classifiers. The strategies used to foresee comprises filtering segmentation and feature extraction to the model. Algorithm: K-Nearest Neighbor, Naïve Bayes.

Proposed system: This research used clinical datasets of 400 patients to determine kidney disease by applying nine machine learning (ML) algorithms. authors identified relevant features from the raw dataset by preprocessing and then predicted chronic kidney disease (CKD) using ML techniques.This study would allow for prompt and accurate treatment of the risk factors identified throughout any appropriate and safe diagnosis of CKD. The authors also compared their results with previously published findings.The nine ML algorithms used in the study were all effective at predicting CKD. The LightGBM algorithm achieved the highest accuracy of 99.00%.The authors identified several relevant features for predicting CKD, including age, gender, blood pressure, and serum creatinine level.The results of the study were consistent with previously published findings[20].

Advantages Of Proposed System: The authors proposed nine machine learning (ML) approaches to predict chronic kidney disease (CKD). The dataset used for this study consisted of 400 records with 14 attributes.
The LightGBM model achieved the highest accuracy of 99%.
The authors concluded that LightGBM is a promising ML approach for predicting CKD.
Algorithm K-nearest neighbors (KNN), support vector machines (SVM), logistic regression (LR), Naïve Bayes, Extra tree classifiers, AdaBoost, Xgboost, and LightGBM.

IV. METHODOLOGY

In the suggested methodology, the process of data pre-processing starts only after the data collection phase is complete. The study employs various classifiers to investigate Chronic Kidney Disease (CKD) prediction, including XGBoost, Naive Bayes, AdaBoost, ExtraTrees Classifier, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machines (SVM). To train and evaluate the dataset, a hold-out validation approach was utilized. This method involved separating the dataset into training and testing subsets to identify the most effective forecasting technique for CKD[21]. A detailed overview of the proposed methodology is illustrated in the accompanying figure.

A. Dataset Collection

The dataset used for this study, titled "Predicting Chronic Kidney Disease based on Health Records," was sourced from Kaggle [16]. It comprises 400 instances and 24 attributes, which include 23 predictive features and one target class attribute. The dataset contains information on various risk factors for CKD, such as high blood pressure, coronary artery disease, pedal edema, diabetes mellitus, specific gravity, appetite, age, anemia, sugar levels, albumin, red blood cells, pus cells, pus cell bacteria, clumps, blood urea, potassium, serum creatinine, hemoglobin, sodium, white blood cell count, packed cell volume, and red blood cell count. These attributes were used to assess the likelihood of kidney disease.

B. Dataset Pre-Processing

The pre-processing phase involved several key steps to prepare the dataset for analysis. Initially, the attribute names from the public online sources were assigned to the dataset. The pre-processing tasks included handling missing values, where the WEKA function "Replace Missing Values" was used to substitute NAs with the mean values of respective attributes. The original dataset of 400 patient records was then refined to 158 instances, consisting of 9 objects, 1 integer, and 14 decimal attributes. This stage also involved data cleaning, feature extraction, and the transformation of categorical variables.

C. Validation Process

Choosing the right validation method is crucial for evaluating the effectiveness of machine learning models. For this study, hold-out validation was selected as it is well-suited for large datasets and provides reliable results [17]. In this method, 30% of the dataset was used for testing, while the remaining 70% was employed for training the models. The performance of each machine learning algorithm was assessed based on metrics such as precision, recall, and F1-Score. The results, including detailed performance metrics and output graphs, are discussed in the result analysis section [19]. A step-by-step flowchart summarizing the entire research process is also provided.

V. CONCLUSION

Chronic Kidney Disease (CKD) poses a significant challenge in the medical field. The impact on patients suffering from CKD can be significantly mitigated if the disease is detected at an earlier stage before it becomes severe. In our study, we utilized a dataset collected from patients over a two-month period to investigate methods for predicting CKD. Our analysis demonstrated a remarkable 99.00% accuracy in detecting and predicting CKD at an early stage. To achieve this, we employed several machine learning algorithms including XGBoost, Random Forest, ExtraTrees Classifier, Naïve Bayes, Logistic Regression, SVC, AdaBoost, LightGBM, and KNN. Among these, LightGBM was found to be the most effective model, outperforming others in terms of accuracy, precision, and F1 score. The outcomes of this research suggest that machine learning-based models hold great promise for developing resources and implementing public health initiatives focused on early CKD detection and patient monitoring. Future work will aim to apply additional datasets from diverse populations and explore further classification techniques such as Deep Learning to enhance the results.

VI. REFERENCES

[1] Gwozdzinski, Krzysztof, Anna Pieniazek, and Lukasz Gwozdzinski. "Reactive oxygen species and their involvement in red blood cell damage in chronic kidney disease." Oxidative medicine and cellular longevity 2021 (2021): 1-19.

[2] Saikat, Abu Saim Mohammad, Ranjit Chandra Das, and Madhab Chandra Das. "Computational Approaches for Structure-Based Molecular Characterization and Functional Annotation of the Fusion Protein of Nipahhenipavirus." Chemistry Proceedings 12.1 (2022): 32.

[3] Saikat, Abu Saim Mohammad, et al. "In-Silico Approaches for Molecular Characterization and Structure-Based Functional Annotation of the Matrix Protein from Nipahhenipavirus." Chemistry Proceedings 12.1 (2022): 21. [

4] R. K. Al-Ishaq, P. Kubatka, M. Brozmanova, K. Gazdikova, M. Caprnda, and D. Büsselberg, “Health implication of vitamin D on the Heidarian, Esfandiar, and Ali Nouri. "Hepatoprotective effects of silymarin against diclofenac-induced liver toxicity in male rats based on biochemical parameters and histological study." Archives of Physiology and Biochemistry 127.2 (2021): 112-118.

[5] “Preventing chronic diseases : a vital investment : WHO global report.” https://apps.who.int/iris/handle/10665/43314?fbclid=IwAR0Fy2HvtoTvEI9cJLm1w8eWXwbKVs0S_UxvWTFenQ60lbjq9VfatepLCiQ (accessed Jan. 23, 2023).

[6] Bastos, Marcus Gomes, and Gianna Mastroianni Kirsztajn. "Chronic kidney disease: importance of early diagnosis, immediate referral and structured interdisciplinary approach to improve outcomes in patients not yet on dialysis." Brazilian Journal of Nephrology 33 (2011): 93- 108.

[7] Roth, Jan A., et al. "Introduction to machine learning in digital healthcare epidemiology." Infection Control & Hospital Epidemiology 39.12 (2018): 1457-1462.

[8] Gopika, S., & Vanitha, M. (2017). Survey on Prediction of Kidney Disease by using Data Mining Techniques. International Journal of Advanced Research in Computer and Communication Engineering, 6(1).

[9] Almasoud, Marwa, and Tomas E. Ward. "Detection of chronic kidney disease using machine learning algorithms with least number of predictors." International Journal of Soft Computing and Its Applications 10.8 (2019).

[10] Vijayarani, S., S. Dhayanand, and M. Phil. "Kidney disease prediction using SVM and ANN algorithms." International Journal of Computing and Business Research (IJCBR) 6.2 (2015): 1-12.

[11] 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 2018.

[12] 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, 2020. 1270 Authorized licensed use limited to: Florida State University. Downloaded on April 19,2023 at 16:52:17 UTC from IEEE Xplore. Restrictions apply.

[13] Chiu, Ruey Kei, et al. "Intelligent systems developed for the early detection of chronic kidney disease." Advances in Artificial Neural Systems 2013 (2013).

[14] Yashfi, ShanilaYunus, et al. "Risk prediction of chronic kidney disease using machine learning algorithms." 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2020.

[15] Wibawa, Made Satria, I. Made DendiMaysanjaya, and I. Made AgusWirahadi Putra. "Boosted classifier and features selection for enhancing chronic kidney disease diagnose." 2017 5th international conference on cyber and IT service management (CITSM). IEEE, 2017.

[16] “Predicting Chronic Kidney Disease | Kaggle.” https://www.kaggle.com/code/csyhuang/predicting-chronic-kidneydisease/notebook (accessed Jan. 23, 2023).

[17] Mamun, Muntasir, et al. "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?." 2022 IEEE World AI IoT Congress (AIIoT). IEEE, 2022.

[18] Mamun, Muntasir, et al. "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis." 2022 IEEE World AI IoT Congress (AIIoT). IEEE, 2022.

[19] Farjana, Afia, and Aaisha Makkar. "Federated Learning for Lung Sound Analysis." Recent Trends in Image Processing and Pattern Recognition: 5th International Conference, RTIP2R 2022, Kingsville, TX, USA, December 1-2, 2022, Revised Selected Papers. Cham: Springer Nature Switzerland, 2023.

[20] Uddin, Md Milon, et al. "Mental Health Analysis in Tech Workplace."

[21] R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen, R. Hasan, “Heart Disease Detection Using ML”, 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), 2023, (Preprint).