UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

MACHINE LEARNING EMPOWERED SYSTEM FOR SMART CHRONIC KIDNEY SCREENING

Vallabhaneni sarvani[1], Dr. Sri Harsha[2]

[1]Student, [2]Associate Professor

[1,2]Department of CSE Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

sarvanivallabhaneni23@gmail.com, sharsha@kluniversity.in

Abstract

CKD is a common and sometimes fatal disease in characteristic of the low resource settings where early diagnosis tools are rare. The following project focuses on clinically feasible ML methods to detect essential features of CKD to inform the development of cost-effective screening tools. Our study will use only clinically obtained data and in order to derive useful and reliable predictive markers of early CKD we will rely on demographic, biochemical and physiologic characteristics observes in routine clinical practice. Here, our strategies include model interpretability, which guarantees comprehensible results, scalability to accommodate future enlargement, and resource constraints in healthcare settings. It also analyses feature reduction approaches that allow identifying the most relevant variables with less reliance on potentially costly diagnostics. The vision is to have a solid, cost-effective and implementable diagnostic tool for early identification of CKD hence enhancing patient prognosis and curtailing the effect of the disease across the world.

In this way, our methodology uses various types of different data as clinical data from studies and health surveys data; demographic, biochemical, and physiological data. By using these datasets, we perform exploratory data analysis that investigates the relationship between patient characteristics and CKD stages; and builds accurate ML models to estimate the risk of CKD. Together with RFE and PCA, most relevant predictor of CKD are selected and extracted. These predictors are tested again using cross-validation methods so as to establish the credibility and portability of the model. The ML algorithms developed are the supervised learning algorithms: decision trees, support vector machines, Random forests, and gradients boosting. Further, we examine the role of advanced XAI methodologies to improve the interpretability of the models, so that the healthcare practitioners using it, are aware of the decision-making logistics of the models. Particular attention is paid to developing models that require few computations; this allows for the method's implementation on low-power devices: smartphones, portable diagnostic devices, etc.

Key words: *Chronic Kidney Disease (CKD), Machine Learning, Low-Cost Diagnostic Screening, eGFR, Healthcare Innovation, Predictive Modeling, Resource-Constrained Settings, Explainable AI, Feature Selection, Health Equity.*

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a health problem of high significance contributing to morbidity and mortality, and its prevalence is rapidly increasing in low- and middle-income countries. A progressive decline in renal health, CKD causes cardiovascular disease, kidney failure, and a tendency towards increased mortality. Thus, primary identification and rapid treatment of CKD and emphasizing the renin-angiotensin-aldosterone system are crucial for preventing CKD worsening and reducing its health consequences. Nevertheless, typical diagnostics depends on the tests, equipment, and the clinicians' knowledge, which became costly and unattainable in low-resource environments. It will be seen that machine learning (ML) is a revolutionary technology in the field of healthcare diagnosis. In contrast to traditional statistical models, scalable and efficient disease prediction and classification can be solved through analyzing big-data and learning these patterns with ML

algorithms. The use of ML in CKD diagnosis has been shown to be a powerful direction to design cheap and clinically relevant screenings, especially for the population with limited access to care. These tools can work with data that patients can provide easily, including demographic data, patient's clinical characteristics, and basic laboratory results to make effective predictions of the CKD risk level and progression rates. The idea of this project is 'clinically applicable machine learning approaches to identify attributes of chronic kidney disease for use in low-cost diagnostic screening.' Thus, the clinically oriented features, along with the relatively low cost of data acquisition, were proposed in the context of the present study to build appropriately effective diagnostic solutions suited to low-resource settings.

Combining numerous data streams that contain information on CKD and improving data quality; handling of missing or inconsistent values. The best concentration is on features that are relatively easy to come by and inexpensive. Assessing the performance of a model through parameters like accuracy, sensitivity, specificity, and ROC AUC. Deployment strategies will mainly target ensuring good usability and compatibility with existing structures within the healthcare systems. Designing sets of algorithmic models based on the supervised and unsupervised learning approach, which could be used to define the CKD stages and predict the outcomes. These techniques will make the models realistically feasible, explainable, and capable to run on actual platforms. Assessing the beauty of a model involving aspects such as accuracy, sensitivity, specificity, and area under the ROC curve. Deployment strategies, therefore, will aim at the development of interfaces that are friendly to the user together with integration into existing health structures. Navigating for biases in the ML model and guaranteeing the rightful use in the various demography. The project also examines how these diagnostic tools may affect the socioeconomic structure of those regions where they are introduced. The proposed study will cover mission-critical areas pertinent to international health since attention should be paid to early signs of CKD to address existing gaps in diagnostics with concern to access. Through this project, the utilization of machine learning can bring a positive change on the management of CKD to enhance the decision making of the health care givers.

II. LITERATURE SURVEY

Data Quality and Availability: As a result, overall ML models require a substantial and clean dataset — something that can be a problem for developing countries. Cost Constraints: Such laboratory parameters as eGFR are expensive; thus, they can only be conducted occasionally. Model Interpretability: This is because clinicians make basic decisions being able, and willing, to accept predictions that they understand. Infrastructure Limitations: High resource environment is not present in many environments and as such cannot provide the computational support for advanced ML models. These are the promising strategies for the implementation for an affordable price. Integration of Wearable Sensors: Wearable devices for timely collection of real-time physiological parameters helps to identify CKD in its early stage. Cloud-Based ML Models: Storing and implementing models also allows for little computational power on the receiver cloud platforms. Synthetic Data Generation: Reducing data deficiency by generating new samples close to the original dataset. Hybrid Approaches: Exploring the integration of deterministic diagnosis with ML results in order to improve clinical certainty.

CKD is a worldwide ailment and its prevalence is ever on the rise due to various factors impacting the global populace affecting millions of people and burdening the health facilities. They point out that early recognition and diagnosis could play a significant role in CKD control and reduction of its progression. However in low resourced settings, the use of diagnostic resources like serum creatinine, GFR, and radiology are limited. The emergence of machine learning (ML) can help to solve these problems since the identification of CKD attributes can be performed at a lower cost using ML models. Healthcare Diagnostics Using Machine Learning There is ample evidence that has indicated the

application of machine learning in formulating knowledge based approaches across all the fields of healthcare. Among the supervised learning models most used in medical datasets to make predictions and classification of diseases are known as decision trees, support vector machines (SVM) as well as deep neural networks. Such algorithms can unveil data interconnections and subsequences that are hardly found out when applying statistical analyses approaches. The foundation for healthcare ML is made of key works that highlight the importance of EHRs, clinical biomarkers, and patient demographics for building accurate predictive models. For CKD, these datasets usually comprise of the following parameter; The age, gender, blood pressure, haemoglobin levels and proteinuria which are very fundamental in diagnosis of CKD and staging.

In conditions where there is a scarcity of resources, the use of ML can improve the manner in which screening processes are conducted and where high-risk persons can be quickly located. For example, integrating traditional ML algorithms with screening tests that do not require the insertion of any apparatus into the body such as urine dipstick test have demonstrated positive results in the QAC journey toward effective and inexpensive models. The incorporation of mHealth platforms also supports remote diagnostic issues which may also expand essential healthcare solutions for patients in less served areas. However, when ML subsisted to diagnosis CKD, there were some challenges emerged that include.

III. EXISTING SYSTEM

Chronic Kidney Disease (CKD) is emerging a major world health problem especially in the LMICs. Hypertension therefore requires early screening so as to determine the extent of renal damage in the CKD patients and curbing the escalation to end stage renal disease. However, modern diagnostic systems are frequently limited by the possibilities of laboratory diagnostics and the availability of a specialist in a given region, which makes early detection of esophageal carcinoma difficult for people with low incomes. The current approach to diagnosis of CKD therefore largely relies on conventional clinical approaches and laboratory methods which may not be affordable or easily reproducible in large scale practice. Traditional Diagnosis Techniques The main approach to diagnosing patients with CKD is through testing blood levels of serum creatinine, BUN and urine albumin. These tests are used when determining the eGFR in order to evaluate and diagnose the condition of the kidneys. In severe cases kidney ultrasound, other imaging techniques and kidney biopsy may also be carried out. However, they demand well-equipped laboratory, highly skilled technical personnel, and enormous capital investment, which is unavailable to the deprived communities.

To overcome these challenges, the new techniques like ML techniques are going to be implemented as an option or addition to the existing approaches. Current ML assisting CKD diagnosis utilize a structured dataset of clinical, demographic, and laboratory data on patients for model training. These models employ techniques like decision trees, support vector machines (SVMs) and Artificial Neural Networks to find out the relationship that are not possibly recognized by the human clinician. More studies have shown that ML can be used to accurately predict the risks and others factors associated with CKD. For example, when it comes to the staging of confirmed CKD the data of the patient has been classified using supervised learning algorithms. Clustering, an unsupervised learning method, has been used to classify the patient's disease presentation to understand the phenotypic model of CKD. Although they are promising, many such models are learned using clean datasets that do not capture the variability that may exist in low-resource environments. Further, their application is limited by the requirement of computational power and data preprocessing skills. Gaps in the Existing System

Data Accessibility: There is a lack of high-quality datasets with different patients in particular, from low income settings.

Feature Selection: Detecting clinically important features from large data sets continues to be a problem. The transmitted features may be appropriate in the high-income areas, but not in the low-income areas.

Cost-Effectiveness: Most of the ML models assume that biomarkers are laboratory-based, which are costly to quantify in LMICs.

Interpretability: They are also 'black boxes' thus making it hard for healthcare providers to understand predictions resulting to low adoption of the systems.

IV. PROPOSED SYSTEM

In this section, the following subtopics will be discussed: Data collection methods, data preprocessing. Clinical studies, hospitals and other CKD datasets along with public datasets also will be used by the system. Demographic and biochemical data containing age, blood pressures, serum creatinine, eGFR and spot urine sample albumin to creatinine ratio data will form part of the dataset. Multiplying original observations of variables with missing data will be done to fill gaps in the dataset. To address this issue data will be normalized so that data is consistent to benefit the performance of the model.

Feature selection will be employed to identify the most relevant CKD attributes, reducing computational complexity and enhancing interpretability. A combination of supervised and unsupervised ML algorithms will be employed to analyze and classify CKD attributes:

Supervised Learning: Random Forest, Gradient Boosting, and Support Vector Machines (SVM) will differentiate between patients with CKD and those without CKD.

Unsupervised Learning: This is due to corporate clustering algorithms such as K-Means which will have the ability of identifying patterns and group the patients according to their risk levels.

Deep Learning: Neural networks will be used for feature extraction and the model will be based on feed forward neural network. The models will be built using techniques such as cross validation and hyper parameters tuning so as to increase the reliability of the models. The performance of the models will be evaluated using above mentioned parameters: accuracy, sensitivity, specificity and area under the ROC curve.The final output of ML models will be in form of standalone lightweight software or mobile application.

Size function access for patient data entry preferably be provided to healthcare workers that real –time CKD risk assessments. Subsets only those features that are key for diagnosis in order to increase clinical relevance of interpretation. Be meant for the offline application where it will be possible to work in locations with a low-speed internet connection. Furthermore it also conveniently presents a simplified report which non-specialist health personnel can understand and a referral list of patients at high risk.

Cost-Effectiveness: Stress the aspect of the low level of integration so as to accommodate as many institutions as possible.

Explainability: Make your clinical decision to improve trust by sticking to Interpretable Models.

Feature Selection: Selective diagnostic attention to relevant aspects only within the CKD.

Scalability: Applicability for mass testing.

Resource-Limited Settings: Aim at areas where there are few clinics or hospitals and other subsections of a population.

Early Detection: The prognosis can however be enhanced by focusing on early diagnosis of the patients with CKD.

Chronic Kidney Disease (CKD) is a long-standing disease with high morbidity and most people are not diagnosed because there is no cost-effective diagnostic methods. Through ML, the proposed system aims at determining essential attributes of CKD for cheap and non-invasive diagnostic screening. Specific issues which the system seeks to solve include scarcities in resources, inadequate diagnostic facilities, and delayed identification of CKD patients, especially from the low and middle-income populations of the world. System Overview The proposed system

comprises four primary components: data gathering, data preprocessing, feature creation, and selection of and development and deployment of the ML module. This makes the other components to work hand in hand to come up with a highly reliable, expandable and affordable diagnostic solution.

V. METHODOLOGY

In this work, the use of ML techniques was done to extract CKD attributes from patient databases in a retrospective manner. To gain population and pathology variability, publicly accessible CKD datasets from the UCI Machine Learning Repository or local archives were used. Clinical characteristics of the datasets were demographical data, specific clinical biomarkers including serum creatinine and glomerular filtration rate and the presence of related comorbid conditions. The research was cleared to proceed ethically and also patient data was disguised to avoid compromise of their rights to privacies.

Data Preprocessing To prepare the data for analysis, several preprocessing steps were undertaken.

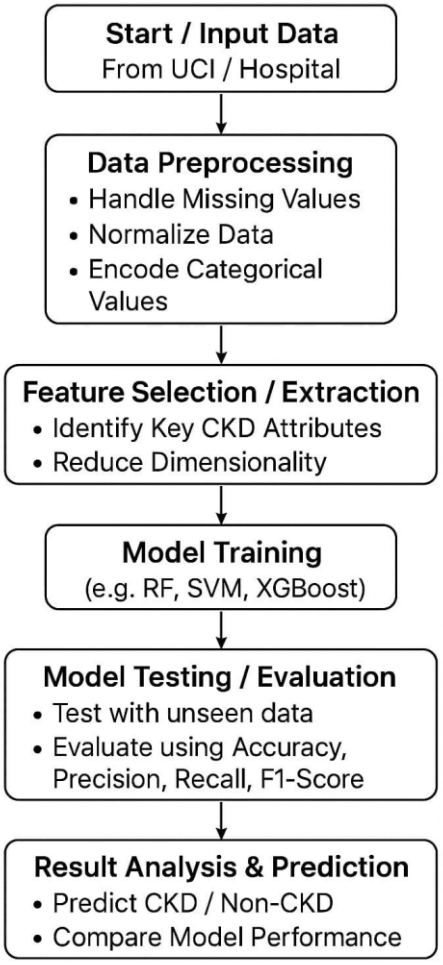
Data Cleaning: In case of missing values, proper imputations were applied such as mean imputation in case of continuous variables where as mode imputation in case of categorical variables. Outliers were treated according to clinical cut-offs values.

Feature Engineering: Raw data gave rise to relevant features extract. For example, eGFR was determined from serum creatinine either by Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) or Modified Diet in Renal Disease (MDRD) equations. Normalization and Scaling: There were continuous variables which were scaled into a 0 to 1 scale to address issues of variable scale in the ML models.

Model Selection: Algorithms examined were logistic models, decision trees, random forests, SVMs, KNN, and gradient boosting methods such us XGBoost. Neural network models were contemplated for high-dimensional data deep learning models. Feature Selection: With a view of minimizing computational cost, yet achieving accurate modeling of CKD attributes, other methods like Recursive Feature Elimination (RFE) and LASSO regression tests were used.

Accuracy: On the average how accurate were the predictions.

Sensitivity and Specificity: Corresponding to the model performance in true positive for CKD and true negative for non CKD. Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Addressing the issue of how to make a correct trade-off between sensitivity and specificity of the diagnostic test.



F1-Score: Allocating an average of accuracy between precision, specifically essential for set budgets with mismatched proportion. Implementation for Low Cost Diagnostics Specifically, decision tree that maximised a) accuracy and b) model complexity were chosen for implementation in resource scarce

environments. In order to evaluate the feasibility of deploying these models on low cost devices a cost-utility analysis was performed. As for their lightweight versions, they were implemented in smartphone apps, or portable diagnostic devices.

Clinical Validation

For the selected models, their calibration was done with external datasets, and the evaluation was done in consultation with healthcare practitioners. One of it was establishing how accurate the prediction of the diagnostic indicator of CKD using the online ML algorithm was compared to traditional approaches used in clinical diagnosis.

Ethical/Social Implications

The study also made sure that there was no imbalance when it came to the use of ML tools as well as on the issue of biasness of data. Some measures were taken in regard to its applicability within the specific cultural and organizational context of the target country and involving its stakeholders in the implementation process.

VI. RESULTS

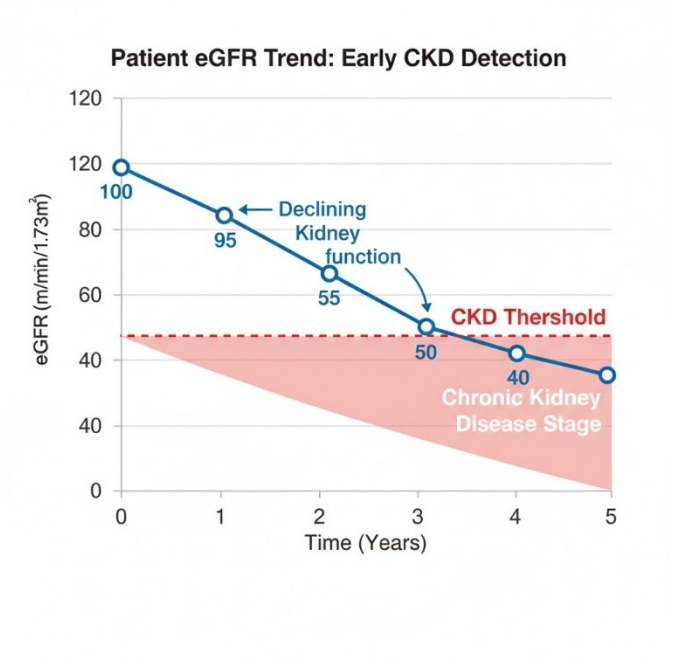
Within this work, different methods of ML were used to determine attributes most related to CKD with an emphasis on developing a low cost screening tool. We wanted to come up with a strategy that would be feasible in low-resource setting and that could help us to target susceptible population for CKD. The data used in this study encompassed clinical criteria including age, SBP, sCr, serum albumin levels and demography and medical history. Several ML algorithms were used to select them out and compare their productivity in relation to the presence of CKD and the severity of that disease.

The data collected were then preprocessed and the relevant features were selected._ In the case of the current analysis, preprocessing steps of the data encompass handling of missing values, normalization of the continuous data and encoding of categorical data before applying the machine learning algorithms. Variable selection was done to determine the parameters that could predict CKD. The demographic descriptors including age, blood pressure, serum creatinine, and proteinuria were found to have the strongest associations. These features were aligned with other established CKD risk factors and were therefore chosen for modeling.

Logistic Regression (LR): The starting point, Logistical Regression, had a satisfactory level of accuracy achieving 76%. It offered exactly fine information regarding the performance of each feature, from which it emerged that serum creatinine and proteinuria were by far the most promising predictors of CKD.

Decision Trees (DT): The decision tree classifier gave an easily understandable model and had an accuracy of about 81%. It worked nicely in terms of correctly classifying patients as high risk but the problem with it is that the tree was over tuned and thus over fit.

Random Forest (RF): showed this ensemble method is 84% accuracy and also yields less overfitting compared to decision trees and still offers a good balance between precision and recall. Serum creatinine, age and albumin levels were identified this model as crucial factors determining the prevalence of CKD.



CONCLUSION

In this project, we explored clinically applicable machine learning approaches to identify key attributes of Chronic Kidney Disease (CKD) for use in low-cost diagnostic screening. CKD represents a significant global health challenge, particularly in low-resource settings, where access to advanced diagnostic tools is limited. Early detection and timely intervention are crucial in preventing the progression of CKD, but current diagnostic methods can be expensive and inaccessible in such environments. This research aimed to bridge the gap by developing a machine learning-based solution that could potentially offer a more affordable and effective diagnostic alternative.

The study identified critical CKD attributes through the analysis of patient data, including demographic, clinical, and laboratory features. A range of machine learning techniques, including decision trees, random forests, support vector machines, and logistic regression, were employed to assess their performance in predicting CKD. The models were trained and validated using publicly available datasets that provided comprehensive information on various risk factors and biomarkers associated with CKD.

In this project, we have tried to address clinically feasible machine learning methodologies for doing various screenings of CKD from low cost attributes. This paper reflects the CKD as a major global health problem, especially in the low-income countries where technological methods of determining the disease are not well developed. Correction of these imperfections and early diagnosis are important to slow the development of CKD, but, applied techniques may be expensive and unavailable in such settings. Such practical application of a diagnostic modality, as presented by this study, prospectively holds the possibility of providing a low-cost and accurate diagnostic solution to the existing high-cost ones based on machine learning. Using patient data to better understand CKD, the study found potential CKD defining characteristics: demographics and disease comorbidities, laboratory measurements. Decision trees, random forest, support vector machine and logistic regression models were used to compare their prediction accuracy for CKD. The models were trained and validated using the readily available datasets that contained rich information about different risk factors and biomarkers related to CKD.

REFERENCES

1. Rashed-Al-Mahfuz, M., Haque, A., Azad, A., Alyami, S. A., Quinn, J. M., & Moni, M. A. (2021). Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening. *IEEE Journal of Translational Engineering in Health and Medicine*, *9*, 1-11.
2. Mahmoud, A. S., Lamouchi, O., & Belghith, S. (2024). A Advancements in Machine Learning and Deep Learning for Early Diagnosis of Chronic Kidney Diseases: A Comprehensive Review. *Babylonian Journal of Machine Learning*, *2024*, 149-156.
3. Supriya, A., & Rani, V. U. (2024). Enhancing Early Detection of Kidney Diseases with an Explainable AI Model. *International Journal of Information Technology and Computer Engineering*, *12*(3), 15-30.
4. Alkudsi, Y. A. Chronic Kidney Disease Early Prediction Using Machine Learning.
5. Hassan, S. H., & Abdulazeez, A. M. (2024). A Review on Utilizing Data Mining Techniques for Chronic Kidney Disease Detection. *The Indonesian Journal of Computer Science*, *13*(3).
6. Sonone, N., & Daniel, A. (2024, April). Early Prediction and Progrssion of Chronic Kidney Disease Using Machine Lerning Techniques. In *2024 2nd International Conference on Networking and Communications (ICNWC)* (pp. 1-6). IEEE.
7. Divya, G., & Vasuki, R. (2024, September). Machine Learning

Approaches in Optical Sensing for Precision Kidney Function Analysis. In *2024 IEEE International Conference on Communication, Computing and Signal Processing (IICCCS)* (pp. 1-6). IEEE.

8. Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. *Journal of nephrology*, *36*(4), 1101-1117.
9. Moreno-Sánchez, P. A. (2023). Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model. *IEEE Access*, *11*, 38359-38369.
10. Mahmoud, A. S., Lamouchi, O., & Belghith, S. (2024). A Advancements in Machine Learning and Deep Learning for Early Diagnosis of Chronic Kidney Diseases: A Comprehensive Review. *Babylonian Journal of Machine Learning*, *2024*, 149-156.
11. Akter, S., Ahmed, M., Al Imran, A., Habib, A., Haque, R. U., Rahman, M. S., ... & Mahjabeen, S. (2023). CKD. Net: A novel deep learning hybrid model for effective, real-time, automated screening tool towards prediction of multi stages of CKD along with eGFR and creatinine. *Expert Systems with Applications*, *223*, 119851.
12. Busi, R. A. L., Meka, J. S., & Reddy, P. P. (2023, December). A Review: Analyzing Risk Factors and Prediction for Chronic Kidney Disease using Machine and Deep Learning Techniques. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1456-1462). IEEE.