Patient-Specific Machine Learning-Based Models for Chronic Kidney Disease Prediction and Treatment in Sri Lankan Population

Wishalya Tissera
Faculty of Computing
Sri Lanka Institute of Information and Technology
Sri Lanka
wishalya.t@sliit.lk

Samadhi Rathnyake
Faculty of Computing
Sri Lanka Institute of Information and Technology
Sri Lanka
samadhi.r@sliit.lk

Samarawila D.R.N.
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka
it20235260@my.sliit.lk

Isurika W.B.M.A.
Faculty of Computing
Sri Lanka Institute of Information and Technology
Sri Lanka
it20785192@my.sliit.lk

Perera J.P.M.L.
Faculty of Computing
Sri Lanka Institute of Information and Technology
Sri Lanka
it20226596@my.sliit.lk

Marasinghe M.M.K.L.
Faculty of Computing
Sri Lanka Institute of Information and Technology
Sri Lanka
it20154226@my.sliit.lk

***Abstract*—** **This undergraduate research project aims to develop and evaluate machine learning models specific to the clinical data and features of Sri Lankan patients to predict the risk of kidney illness. The project will also create a system for making treatment recommendations based on the estimated risk and patient-specific information to offer personalized treatment options. The results might lead to more precise kidney disease diagnosis and treatment, which would eventually reduce the effect on the healthcare system and improve patient outcomes. The project will utilize existing datasets and conduct surveys and interviews with healthcare providers and patients to gather insights into the factors that may influence the models' performance. The findings will be presented in a research report that may guide evidence-based policies and practices to address kidney disease in Sri Lanka. This high-level research study addresses the research gap regarding patient-specific ML models for renal disease prediction and treatment in Sri Lanka, which makes an important contribution to healthcare. The outcomes may improve patient outcomes, effectiveness of therapy, and diagnostic accuracy. In the end, this study intends to reduce the financial burden of kidney disease on Sri Lanka's healthcare system and set a standard for such programs across the world.**

Keywords: kidney disease, machine learning, prediction models, patient-specific, Sri Lanka, early detection, treatment, healthcare providers, patients, diagnosis, burden, treatment recommendation system.

I. Introduction

With an approximate prevalence of around 16%, kidney disease is still a serious public health problem in Sri Lanka and affects a sizeable section of the population. To stop future deterioration and maintain general health, early detection and proper care of kidney disease are important. In Sri Lanka, medical professionals have historically used patient interviews to determine the possibility of kidney disease by evaluating symptoms and risk factors. The exploration of cutting-edge technology, however, is gaining popularity to improve the accuracy of kidney disease prediction.

Doctors often use the traditional approach, which involves gathering the patient's medical history and asking about symptoms such leg swelling fatigue, changes in urine flow, and difficulty urinating. Along with lifestyle choices like drinking or smoking, risk factors including medical history and medication use are also considered because they might raise someone's chance of developing kidney disease. Based on these replies, medical professionals decide if more diagnostic procedures, such as blood and urine testing, are necessary to confirm the diagnosis.

Even while machine learning (ML) algorithms have shown significant potential in seeing patterns in huge datasets, applying them to reliably predict a patient's risk of developing kidney disease is still difficult. To provide effective preventative measures and treatments, high levels of accuracy must be achieved. Therefore, it is crucial to investigate cutting-edge strategies to improve ML algorithms' prediction abilities in the context of kidney diseases.

This research project aims to create a smartphone application that uses ML algorithms to precisely forecast kidney disease in the Sri Lankan population in answer to this problem. This application might revolutionize kidney disease identification and treatment by utilizing technology to give a readily accessible tool for individuals and medical professionals.

The goal of this study is to overcome the shortcomings of current prediction techniques and increase the precision of kidney disease diagnosis. Patients may simply estimate their risk of kidney disease by using a simple and user-friendly smartphone application. The program may be used by healthcare practitioners as a helpful decision-support tool, enabling early detection and individualized treatment plans.

This study aims to improve the mobile application's predictive skills, assuring trustworthy and accurate forecasts of kidney diseases, through the deployment of cutting-edge ML algorithms and the integration of patient-specific data. By increasing the rates of early identification, enabling prompt treatments, and eventually improving patient outcomes, the findings of this study have the potential to have a considerable influence on Sri Lankan healthcare practices.

This project aims to change kidney disease prediction in Sri Lanka by embracing technology and integrating machine learning algorithms. The methodology, data collecting, analysis, and assessment methods used to create and evaluate the mobile application are described in depth in the following sections. The findings of this study could significantly affect how kidney disease is managed, advancing medical procedures in Sri Lanka.

II. Literature Review

Elias Dritsas and Maria Trigka, researchers affiliated with the University of Patras in Greece, studied to predict chronic kidney disease (CKD) using machine learning (ML) techniques. CKD is a condition that causes a gradual decline in kidney function and can ultimately lead to end-stage renal disease and death if not correctly diagnosed and treated. The researchers utilized a class-balancing approach to address the non-uniform distribution of instances in the two classes, performed feature ranking and analysis, and trained and evaluated several ML models based on various performance metrics. The study's results demonstrated that the Rotation Forest (RotF) model outperformed the other models, achieving an Area Under the Curve (AUC) of 100% and high levels of precision, recall, F-measure, and accuracy, with a rate of 99.2% [1].

Chin-Chuan Shih and the team conducted a research study in Taiwan in 2020 to develop efficient risk prediction models for complications and mortality rates associated with chronic kidney disease (CKD). To predict early CKD, the study utilized four data mining algorithms, which included a classification and regression tree, a C4.5 decision tree, a linear discriminant analysis, and an extreme learning machine. The study collected data from an adult health examination program conducted between 2015 and 2019, including information from 19,270 patients from 32 chain clinics and three special physical examination centers. The predictive variable used was the glomerular filtration rate (GFR), and 11 independent variables were considered. Out of the four models, the C4.5 decision tree algorithm demonstrated superior performance, with higher accuracy, sensitivity, specificity, and area under the curve metrics. The study identified several significant risk factors for early CKD, including Urine protein and creatinine ratio (UPCR), Proteinuria (PRO), Red blood cells (RBC), Glucose Fasting (GLU), Triglycerides (TG), Total Cholesterol (T-CHO), age, and gender. The proposed risk prediction models could assist in predicting early CKD and provide valuable insights into managing this healthcare priority by considering personality and health examination representations [2].

In 2023, Piyawat Kantagowit, Fangyue Chen, Tanawin Nopsopon, Arisa Chuklin, and Krit Pongpirul prepared a systematic review protocol to use ML to diagnose CKD. CKD is a significant contributor to morbidity and mortality worldwide, and ML-based decision-support tools have been developed to aid in various aspects of CKD care. The systematic review protocol aims to assess the performance and report the quality of prognostic and diagnostic ML models in CKD diagnosis and prediction. The review will include a systematic search of various databases and compare ML-based models' performance with non-ML-based models as the primary outcome. The secondary analysis will include model use cases, construct, and reporting quality. The results of this systematic review will offer clinicians and technical specialists' valuable insights into the current development and potential standardization of ML in CKD care [3].

In the year 2023, a research study was carried out by Ariful Islam, Ziaul Hasan Majumder, and Alomgeer Hussein to examine the capability of using machine-learning approaches for the early diagnosis of chronic kidney disease (CKD). Early detection and appropriate therapy can slow down or halt the progression of the disease, and the use of machine-learning approaches can aid in achieving an early diagnosis. The study investigated the relationship between data factors and the characteristics of the target class and developed a collection of prediction models using machine learning and predictive analytics. At the outset, 25 different variables were taken into consideration, however, only the top 30% of those parameters were found to be most effective for identifying CKD. Twelve classifiers based on machine learning were tested using a supervised learning environment. The XgBoost classifier emerged as the most efficient, with an accuracy rate of 0.983, precision of 0.98, recall of 0.98, and F1-score of 0.98. The study demonstrates the potential of machine learning and predictive modeling to discover new solutions for early diagnosis of CKD and other diseases. Recent advancements in machine learning have demonstrated promising outcomes in the early detection of kidney disease and beyond [4].

In conclusion, Machine Learning (ML) has shown great potential in predicting and diagnosing chronic kidney disease (CKD), but some limitations still need to be addressed. One of the main limitations of ML-based models in predicting and diagnosing CKD is their accuracy. Although these models have shown promising results, they still need to be improved to achieve higher levels of accuracy. Another area for improvement is the availability of these models for day-to-day users. These models are often complex and require specialized knowledge, limiting their accessibility.

Overall, ML-based models have the potential to revolutionize the diagnosis and treatment of CKD. Still, more work needs to be done to ensure their accuracy and accessibility for day-to-day users. By addressing these limitations, we aim to improve the quality of care for patients.

III. Methodology

This section presents a comprehensive methodology for conducting the research study on chronic kidney disease (CKD) patients in Sri Lanka. The methodology encompasses manual data collection from hospitals, dataset creation, data preprocessing, cross-validation, machine learning model development, integration with a mobile application, and Agile project management. This section provides a detailed description of each step involved in the research process, including relevant mathematical formulas and calculations. Manual Data Collection:

The first step in the research study involved the manual collection of CKD patient data from various hospitals in Sri Lanka. To ensure a representative sample, field visits were conducted to multiple hospitals across different regions. The data collection process focused on gathering information regarding key factors associated with CKD, including gender, presence of diabetes, family history of the disease, obesity, smoking habits, and urinary obstruction. Medical records, patient interviews, and consultations with healthcare professionals were utilized to obtain accurate and reliable data.

A. *Dataset Creation*

Upon collecting the CKD patient data, the next step involved creating datasets for further analysis. The collected data were organized and structured in a suitable format to facilitate effective data processing and modeling. Various attributes such as patient demographics, medical history, and lifestyle factors were incorporated into the dataset. Additionally, relevant metadata, such as data source and timestamp, were included to ensure traceability and transparency.

B. *Data Preprocessing*

Data preprocessing plays a crucial role in preparing the collected dataset for machine learning analysis. This step involved several data cleaning and transformation techniques to enhance data quality and reliability. Missing values were handled using appropriate imputation methods, such as mean imputation or regression-based imputation. Numerical features were normalized or standardized to eliminate scale-dependent biases. Feature scaling was performed to bring all features to a similar range and prevent dominance of certain variables. Outlier detection and removal techniques, such as the Z-score method or the interquartile range (IQR), were applied to mitigate the impact of outliers on model performance.

C. *Sample Size Calculation*

To determine the required sample size, the following formula was employed:

$$n = z^2 * p * (1 - p) / d^2$$

where n represents the sample size, z is the desired level of significance, p is the expected prevalence of CKD, and d is the desired margin of error. This calculation was used to estimate the appropriate number of CKD patients to include in the study.

D. *Data Split and Cross-Validation*

To evaluate the performance and generalization capability of the machine learning models, the dataset was divided into training and testing sets. The commonly used train-test split ratio of 70:30 or 80:20 was employed. Additionally, to assess the model's robustness, cross-validation techniques, such as k-fold cross-validation, were applied.

E. *K-fold Cross-Validation*

For k-fold cross-validation, the dataset was divided into k equal-sized folds. Each fold was then used as a testing set while the remaining folds were utilized for training. This process was repeated k times, with each fold serving as the testing set once. The results were averaged to obtain a more reliable estimation of the model's performance. The value of k was determined based on the research requirements.

F. *Machine Learning Model Development*

Several machine learning models were developed using the preprocessed dataset to predict the likelihood of CKD based on the collected patient factors. Various algorithms, including logistic regression, decision trees, random forests, support vector machines, and ensemble methods, were explored to identify the model with the highest precision, accuracy, and overall performance. Model hyperparameters were tuned using techniques such as grid search or random search to optimize performance.

G. *Feature Scaling*

During the machine learning model development, feature scaling was applied using the formula:

$$x_i = (x_i - mean(x)) / std(x)$$

where x_i represents an individual feature value, mean(x) is the mean of the feature values, and std(x) is the standard deviation of the feature values. This calculation ensured that the numerical features were normalized and had a similar scale, reducing bias, and improving the performance of the machine learning models.

H. *Integration with Mobile Application*

To facilitate broader accessibility and utilization of the research findings, the developed machine learning model was integrated into a mobile application. The application was developed using React Native, a popular framework for building cross-platform mobile applications, and NodeJS for backend development. This integration allowed CKD patients and healthcare professionals to access the predictive model and obtain personalized risk assessments conveniently through their mobile devices. The integration process involved designing and implementing appropriate APIs (Application Programming Interfaces) to enable communication between the mobile application and the machine learning model.

I. *Agile Methodology*

Throughout the research study, Agile methodology was employed as the project management approach. The Agile framework facilitated iterative development, continuous feedback, and collaboration among the research team members. The project was divided into sprints, with each sprint focusing on specific tasks and deliverables. The team conducted regular meetings, including daily stand-ups, sprint planning sessions, and retrospective meetings, to discuss progress, address challenges, and prioritize tasks. This iterative approach allowed for flexibility and adaptability to incorporate changes and enhancements during the research process.

This methodology section outlined a detailed step-by-step approach for conducting the research study on CKD patients. The manual data collection process involved visiting hospitals in Sri Lanka and gathering relevant patient information. The collected data were then used to create datasets, followed by thorough data preprocessing to ensure data quality and suitability for analysis. Cross-validation techniques were employed to assess model performance, and various machine learning algorithms were developed to predict CKD likelihood. The research findings were integrated into a mobile application using React Native and NodeJS technologies, providing convenient access to personalized risk assessments. Throughout the research study, Agile methodology was adopted for efficient project management. By following this methodology, the research aimed to generate valuable insights into factors associated with CKD and develop an accurate predictive model for early detection and prevention of the disease.

IV. Results and Discussion

This section provides an overview of the evaluation conducted for the kidniFy application, focusing on how accurate the prediction model is. We used different measures to see how well the model performs, like precision, recall, F1-score, and accuracy.

To evaluate the model, we set up an experiment with specific hardware and software settings. This helped us create a consistent environment that others can use to replicate our results. The dataset we used was carefully chosen to represent real-life situations. We explain where we got the data from, how we collected it, and the steps we took to make sure it was reliable.

We followed a specific method to evaluate the model's accuracy. This included dividing the data into different parts for training, validation, and testing. We trained the model using the training data and then measured how well it predicted outcomes using the testing data. We also considered any potential biases or limitations in our evaluation and took steps to address them.

Finally, we analyzed the results we obtained. We looked at different measures to see how well the model performed. We highlighted its strengths and weaknesses and compared it to other existing methods in the field. This showed that the kidniFy application is effective and reliable in providing accurate predictions.

V. Conclusion and Future Work

In this research paper, we presented a comprehensive study on the development and evaluation of patient-specific machine learning-based models for chronic kidney disease (CKD) prediction and treatment in the Sri Lankan population. The aim of this study was to address the research gap regarding patient-specific ML models for renal disease prediction and treatment, and ultimately improve patient outcomes, treatment effectiveness, and diagnostic accuracy.

Through the utilization of existing datasets, surveys, and interviews with healthcare providers and patients, we gathered insights into the factors that may influence the performance of the ML models. The manual data collection process involved visiting hospitals in Sri Lanka, and the collected data were carefully structured and preprocessed to ensure data quality and reliability.

We employed various machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and ensemble methods, to develop prediction models for CKD. The models were trained and evaluated using appropriate performance metrics, including precision, recall, F1-score, and accuracy. The evaluation process demonstrated that the developed models, integrated into the kidniFy mobile application, provide accurate predictions and risk assessments for CKD in the Sri Lankan population.

The results of this study have significant implications for healthcare practices in Sri Lanka. The integration of the developed machine learning models into the mobile application allows for personalized risk assessments and early detection of CKD, enabling healthcare professionals to provide timely and targeted treatment plans. The application's user-friendly interface makes it accessible to individuals, empowering them to estimate their risk of kidney disease and take proactive measures to manage their health.

Furthermore, the research findings have the potential to guide evidence-based policies and practices related to kidney disease in Sri Lanka. By improving the accuracy of CKD prediction and treatment, this study aims to reduce the burden on the healthcare system and set a standard for similar programs worldwide.

While this research study has made significant contributions to the field of chronic kidney disease prediction and treatment in the Sri Lankan population, there are several avenues for future work and enhancements:

A. Expansion of the dataset:

To further improve the accuracy and generalizability of the ML models, future work should focus on expanding the dataset by collecting more patient data from diverse regions and demographics. This would allow for a more comprehensive analysis of the factors influencing CKD and enable the development of even more accurate prediction models.

B. Integration of additional data sources:

Incorporating data from wearable devices, electronic health records, and genetic information into the models could provide valuable insights into personalized risk assessments and treatment recommendations. Future work should explore the integration of these additional data sources to enhance the precision and effectiveness of the prediction models.

C. Continuous model refinement:

ML models are iterative in nature, and continuous refinement is necessary to adapt to evolving data patterns and improve prediction accuracy. Future work should focus on regularly updating and fine-tuning the ML models based on new data and insights to ensure the models remain up to date and robust.

D. *Real-world deployment and evaluation:*

Deploying the kidniFy mobile application in real-world clinical settings and conducting extensive evaluation studies would provide valuable feedback on its usability, effectiveness, and impact on patient outcomes. Future work should involve collaborations with healthcare institutions and organizations to conduct large-scale trials and evaluate the application's performance in diverse patient populations.

E. *Integration of explainability techniques:*

Enhancing the interpretability and explainability of the ML models can inspire trust and acceptance among healthcare professionals and patients. Future work should explore the integration of explainability techniques, such as feature importance analysis and model visualization, to provide insights into the decision-making process of the prediction models.

F. *Long-term monitoring and follow-up:*

Tracking patients' health outcomes and treatment responses over an extended period would allow for the assessment of the long-term effectiveness of the prediction models and personalized treatment plans. Future work should focus on establishing long-term monitoring mechanisms and conducting follow-up studies to evaluate the impact of the kidniFy application on patient outcomes and healthcare costs.

References

[1] M. T. Elias Dritsas, "MDPI," 2022. [Online]. Available: https://bit.ly/3KknHjF.

[2] C.-J. L. G.-D. C. C.-C. C. Chin-Chuan Shih, "National Library of Medicine," July 2020. [Online]. Available: https://bit.ly/44RXely. [Accessed April 2023].

[3] F. C. T. N. A. C. P. Piyawat Kantagowit, "PLOS ONE," February 2023. [Online]. Available: https://bit.ly/44Qwwde.

[4] Z. H. M. A. H. Ariful Islam, "Science Direct," 2023. [Online]. Available: https://bit.ly/3QndIOd.