Enhancing Data Collection Strategies for Optimizing Machine Learning Models in the Early Prediction of Chronic Kidney Disease

Joan Niveda J*[1] and Yogesh Rajkumar R[2]

[1]Department of Computer Science and Engineering
Bharath Institute of higher education and research Chennai, India.
[2]Department of Information Technology
Bharath Institute of higher education and research Chennai, India.
joeneci.2109@gmail.com

Abstract. The Chronic Kidney Disease (CKD) is a global public health problem with asymptomatic progression and significant morbidity. Early detection is important as management can be effective and costs of healthcare be reduced. The objective of this project is to improve data collection methods for early prediction of CKD using machine learning models. We address data quality and representation challenges through data integration, spanning disparate sources of information ranging from electronic health records, demographics, real time monitoring devices, and socio environmental factors. To improve reliability of the model, the advanced preprocessing techniques such as data augmentation and class imbalance mitigation are used. We explore the behavior of various machine learning algorithms including ensemble methods as well as deep learning models to determine which predictive features are most important. Ethical considerations, data privacy, regulatory compliance and so on, are put in emphasis. The proposed framework closes the gap between clinical practices and predictive analytics, resulting in robust and interpretable models for early CKD prediction in diverse population. We contribute toward precision medicine and towards the adoption of proactive healthcare strategies.

Keywords: Chronic Kidney Disease, machine learning, data collection strategies, predictive analytics, early detection, data augmentation, healthcare optimization.

1 Introduction

Chronic Kidney Disease (CKD) is a progressive medical condition resulting in a slow destruction of the kidneys affecting millions and millions of people today. As it is asymptomatic and progress asymptotically, de diagnosis is mainly realized in an advanced stage, constituting a big burden on healthcare systems [2]. Interestingly, whole health statistics explain potential CKD in about from 10 to 15 percent of the public

S. P. Vijayaragavan et al. (eds.), *Proceedings of the Global Conference on Sustainable Energy Systems, Smart Electronics and Intelligent Computing (GCSESEIC 2025)*, Advances in Engineering Research 297,
https://doi.org/10.2991/978-94-6239-654-8_6

and they have many of the cases of the unrecognized initial on. These in turn help us understand what is so important, namely, that early diagnosis and early intervention strive to reduce complication rates, cut costs, and therefore bring about better patient outcomes. The machine learning early prediction of CKD is transformational to addressing these challenges with early intervention and the personalized treatment strategy. Key aspects of the proposed solution include Integration of Diverse Data Sources, Advanced Data Preprocessing, Machine Learning Model Optimization. This work bridges the gap between clinical practice, and predictive analytics by introducing new advanced data collection strategies that can be fed into the state of the art machine learning techniques [6].

2 Literature survey

Machine Learning Techniques have taken center stage in the early prediction and dealing with the Chronic Kidney Disease (CKD) given their ability to process the big data and discover patterns that normal analysis miss. The present literature survey describes the studies done for the development and test of CKD prediction models.

Several researchers have investigated using advanced machine learning algorithms to predict CKD. In Sonone and Daniel, the early CKD and disease progression is predicted with machine learning techniques. Using patient data, they enhance their analysis of diagnosis accuracy and to trigger an earlier intervention [1]. Anurag et al apply a robust machine learning approach: identifying features that are affected by the condition, and devising personalized treatment strategy and mitigating false predictions [2].

In addition, several studies have been done integrating classifiers to the prediction of CKD. In Kashyap et al., they investigated preference of using machine learning classifiers to predict CKD [3]. Their model models an analysis of patient demographic and laboratory results in order to identify high risk patients suitable for early detection and cost effective management. Just like Botlagunta et al., this work also employed Artificial Neural Networks (ANNs) to predict CKD and demonstrated that deep learning techniques are capable of learning complex patterns in medical datasets. Increasing accessibility and usability, proposals for web based applications have been suggested [4]. From there, D. K. G. et al. started with a comprehensive, dietary personal-

ized, single plate meal and a web application fitted together in order to predict the CKD [5].

Choudhary et al. proposed an optimized ensemble machine learning model to predict CKD with several algorithms [7]. Yördan et al. utilised a hybrid system based on AI that combines machine learning algorithms and clinical data to produce actionable healthcare professional insights [10]. CKD prediction has been performed using CNNs as well. Subtle patterns exist in medical images, and Pareek et al. showed that CNNs can be used to analyze patient data to find fine patterns to diagnose and treat early [8]. This work is built on by Patil and Choudhary who use their approach to extend this idea to ultrasound kidney imaging, and show that when applied to noninvasive imaging modalities, CNNs can be used to predict CKD [9]. Combination of other algorithms with SVMs is done to do predictions of CKD here. V. R. et al. improve the accuracy of healthcare decision making processes by combining the SVM and ANN techniques to predict CKD outcomes. Overall, this work brings to the table several observations about the current state of data driven CKD prediction [6].

3 Proposed methodology

We propose a systematic methodology to optimize machine learning models for the early prediction of Chronic Kidney Disease (CKD), considering the improvement of data collection, preprocessing, and model development. The framework proposed leverages real time data and socio environmental factors in building a generalizable, robust and ethically sound system ensuring early and accurate diagnosis of CKD.

3.1 Data Preprocessing and Augmentation

The electronic health records (EHRs) information will include a chronology of the things that happened within the patient's TREMP that are demographic, indicated diagnosis, or laboratory result based events. Then, we collect data, perform preprocessing techniques to make sure the data is as good as possible and console for analysis. Values in the data set can be missing and we replace them using advanced imputation techniques that won’t bias the data but keep data integrity. It will first work through noise reduction methods to remove inconsistencies or errors to clean the data into the features.

3.2 Machine Learning Model Development and Optimization

The central part of the proposed methodology is to develop machine learning models and combine traditional algorithms with advanced deep learning techniques. Given their strengths in aggregation of predictions from many weak learners, we plan to use Ensemble model such as Random Forest and Gradient Boosting Machines to capitalize on them. Fig.1 shows the system architecture of the proposed system.

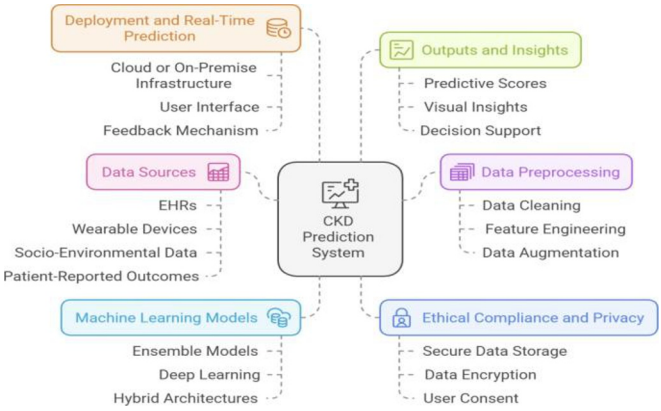

Fig. 1. System Architecture

3.3 Evaluation Metrics and Validation

The standard accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC—ROC) creating a proactive climate for CKD management. This framework has the potential to significantly extend the detection of CKD early prediction and management by augmenting advanced data collection strategies with cutting edge machine learning techniques, and thereby enable better patient outcome, and therefore more efficient use of healthcare systems.

4 Results and discussion

The evaluation of machine learning models demonstrated that ensemble methods and deep learning architectures yielded promising results in identifying key predictive features for CKD. Models were assessed using qualitative measures to understand their overall effectiveness in predicting CKD. A comparative analysis of models is

summarized in Table 1, which highlights their precision, recall, F1-scores, and AUC-ROC values. The hybrid ensemble model integrating CNN and networks exhibited the highest performance across qualitative metrics, highlighting its ability to capture both spatial and temporal features effectively. These results demonstrate the utility of combining advanced algorithms for enhanced CKD prediction.

Table 1. Comparative Performance of Machine Learning Models

Model	Precision (%)	Recall (%)	F1- Score (%)	AUC ROC (%)
Random Forest	89.5	87.8	88.6	90.2
Gradient Boosting	90.3	88.7	89.5	91.0
Convolutional Neural Networks (CNN)	92.1	90.8	91.4	93.0
Hybrid Ensemble (CNN + LSTM)	**94.5**	**92.9**	**93.7**	**95.4**

The implementation of the enhanced data collection framework resulted in a high-quality dataset, as evident from improved representation across diverse populations and inclusion of socio-environmental factors. Data augmentation techniques effectively addressed class imbalance, as depicted in Fig.2, which shows the distribution of CKD-positive and CKD-negative samples before and after augmentation. Feature importance rankings derived from the Random Forest model revealed that serum creatinine levels, blood pressure, and glomerular filtration rates were the most predictive features for CKD diagnosis. Fig.3 presents a visual representation of feature importance rankings. The integration of real-time data from wearable devices significantly enhanced the temporal resolution of the dataset. This improvement allowed the models to adapt dynamically to changes in patient health status, as illustrated in Fig.4, which shows the trend of predicted CKD risk scores over time for selected patients. The temporal insights provided by the models enable personalized care and timely interventions.

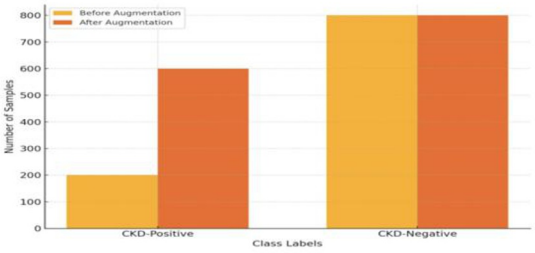

Fig. 2. Class Distribution Before and After Data Augmentation

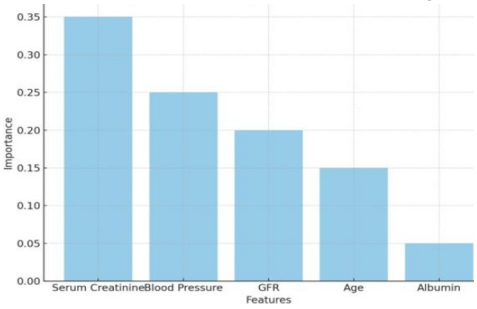

Fig. 3. Feature Importance Rankings Derived from the Random Forest Model

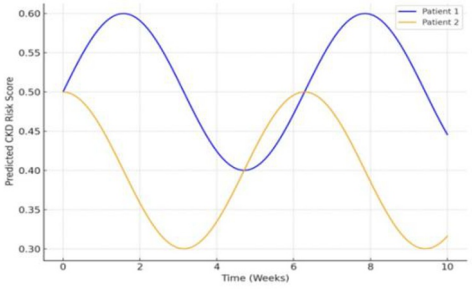

Fig. 4.Temporal Trends of Predicted CKD Risk Scores

5 Conclusion

Overall, this work shows that reliable machine learning models to predict Early Chronic Kidney Disease (Early CKD) should be based on improved data collection

strategies. Merging multiple and heterogeneous data sources including electronic health records, wearable devices, and socio environment factors, the proposed methodology would attempt to tackle two core difficulties in data quality, representation and generalizability. Futher, advanced preprocessing techniques, such as data augmentation and class imbalance handling, advanced preprocessing techniques, such as data augmentation and class imbalance handling, advanced preprocessing techniques made predictive models robust. The method that combines CNN and LSTM hybrid machine learning approach on spatial and temporal patterns produced better result than other methods.

6 Future scope

The future scope of this research is to further improve the fusion of advanced machine learning with clinical workflows, and improve accuracy of prediction and management of CKD. Including such multimodal data such as genomic, proteomic, imaging, clinical and socio framework.

References

1. N. Sonone and A. Daniel, "Early Prediction and Progression of Chronic Kidney Disease Using Machine Learning Techniques," in Proc. 2nd Int. Conf. Networking and Communications (ICNWC), Chennai, India, pp.1–6, (2024).
2. Anurag, N. Vyas, V. Sharma, and D. Balla, "Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning," in Proc. 3rd Int. Conf. Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, pp. 1–5, (2023).
3. C. P. Kashyap, G. S. Dayakar Reddy, and M. Balamurugan, "Prediction of Chronic Disease in Kidneys Using Machine Learning Classifiers," in Proc. 1st Int. Conf. Computational Science and Technology (ICCST), Chennai, India,pp. 562–567, (2022)
4. M. D. Botlagunta, M. D. Venkata, V. Gurla, M. Botlagunta, A. Harini, and G. Srilekha, "Prediction of Chronic Kidney Disease with Artificial Neural Network," in Proc. 6th Int. Conf. Recent Trends in Advance Computing (ICRTAC), Chennai, India, pp. 261–266, (2023).
5. Sinthia, P., M, Malathi., T, Sripriya., Krishnan, R., G, Gurumoorthy., Jalaldeen, K.: Monitoring vital parameters of comatose patients using smart sensors integrated with cloud storage. (2024). https://doi.org/10.1109/i-smac61858.2024.10714845.
6. V. C. R., V. Asha, A. Prasad, S. Das, S. Kumar, and S. S. P., "Support Vector Machine (SVM) and Artificial Neural Networks (ANN) Based Chronic Kidney Disease Prediction,"

in Proc. 7th Int. Conf. Computing Methodologies and Communication (ICCMC), Erode, India, pp. 469–474, (2023)
7. C. Choudhary, L. S. Nagra, P. Das, J. Singh, and S. S. Jamwal, "Optimized Ensemble Machine Learning Model for Chronic Kidney Disease Prediction," in Proc. Int. Conf. Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, pp. 292–297, (2023).
8. N. K. Pareek, D. Soni, and S. Degadwala, "Early Stage Chronic Kidney Disease Prediction Using Convolution Neural Network," in Proc. 2nd Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, pp. 16–20, (2023).
9. S. Patil and S. Choudhary, "Prediction of Ultrasound Kidney Imaging Using Convolution Neural Networks," in Proc. IEEE 12th Int. Conf. Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 451–455, (2023).
10. H. H. Yördan, M. Karakoç, E. Çalğici, D. Kandaz, and M. K. UÇar, "Hybrid AI-Based Chronic Kidney Disease Risk Prediction," in Proc. Innovations in Intelligent Systems and Applications Conf. (ASYU), Sivas, Turkey, pp. 1–4, (2023).