International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v10i10.869
Volume 10, Issue 10, October 2023

Research Paper

Advancing Chronic Kidney Disease Diagnosis: A Predictive Model Using Random Forest Classifier

[1]Polishetty Pranay, [2]Mandadi Rahul Reddy, [3*]K Venkatesh Sharma

[1,2] B.Tech Student ,Department of Computer Science & Engineering, CVR College of Engineering, Rangareddy Dist, Telangana, India
[3]Professor, Department of Computer Science & Engineering, CVR College of Engineering, Rangareddy Dist, Telangana, India

e-mail: pranay070902@gmail.com,rahulreddymandadi@gmail.com, venkateshsharma.cse@gmail.com

**Corresponding Author: venkateshsharma.cse@gmail.com*

Received: 18/09/2023, Revised: 30 /09/2023, Accepted: 19/10/2023 Published: 30/10/2023

Abstract: *In the contemporary medical landscape, Chronic Kidney Disease (CKD) poses substantial challenges, often remaining undetected until severe damage ensues due to current systems' diagnostic limitations. Traditional methods grapple with issues like delayed diagnosis and intensive resource utilization, creating a pressing need for an advanced, efficient approach. Addressing this, our research introduces a ground-breaking predictive model using a Random Forest Classifier, tailored for early CKD detection. We meticulously pre-processed our data, ensuring its reliability, and employed the Random Forest method, known for its precision and ability to manage complex datasets. The model's performance, tested against a comprehensive dataset, achieved an extraordinary accuracy of 95%, highlighting its proficiency in early risk identification and potential in revolutionizing CKD management. This study signifies a remarkable stride in healthcare, offering a precise, scalable, and economical solution for CKD early intervention. By successfully pinpointing CKD onset at initial stages, our model facilitates prompt medical response, enhancing patient prognosis and reducing associated healthcare burdens. Furthermore, it sets the stage for extensive AI integration in diagnostic practices, promising substantial improvements in preventive care and health system efficiency. The implementation of this predictive tool is poised to significantly diminish CKD-related complications and fatalities, emphasizing machine learning's transformative impact in advancing global health standards. This model's integration represents a monumental leap in medical diagnostics, combining innovative technology with profound healthcare implications.*

Keywords: *Diabetes Prediction, Machine Learning, Classification, Ensemble Techniques, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Random Forest, Accuracy.*

---

1. Introduction

In the evolving realm of medical diagnostics, the integration of artificial intelligence and machine learning marks a paradigm shift, significantly altering conventional methodologies. This revolution is particularly impactful in chronic disease management, where early detection is pivotal to improved patient prognosis and cost reduction. Chronic Kidney Disease (CKD), characterized by its insidious onset and often imperceptible progression, epitomizes the diseases that substantially benefit from these advanced predictive techniques.

CKD silently afflicts a global demographic, posing a formidable challenge to healthcare systems due to its late diagnosis, which often precipitates limited therapeutic avenues and increased mortality rates (Li et al., 2020) [1]. In this milieu, machine learning emerges as a beacon of innovation, offering precise, nuanced, and timely disease prediction capabilities (Amiri et al., 2019) [2].

Our research paper, "Chronic Kidney Prediction using Random Forest Classifier," ventures into this innovative domain, proposing a machine learning-based methodology for early CKD detection. Utilizing the Random Forest Classifier's robustness, a model known for its precise predictive capabilities and versatility in handling complex datasets, we aim to establish a dependable early warning system for CKD. This system's inception is a response to the urgent need for more efficient, scalable, and sensitive diagnostic methods, transcending the limitations of traditional approaches like serum creatinine tests and glomerular filtration rate estimation.

Predicting CKD is fraught with complexities, primarily due to the subtle nature of its biological markers and the overshadowing presence of more conspicuous symptoms.

Existing medical protocols for CKD are often reactive, hindered by resource limitations and the disease's asymptomatic nature in its initial stages. These challenges necessitate a more proactive, resource-efficient approach, underscored by recent studies highlighting machine learning's efficacy in chronic disease management (Das & Chakraborty, 2020) [3]. Our research is an endeavor to bridge this gap, harnessing machine learning's predictive acumen to facilitate a scalable and sensitive early CKD detection system, integral to proactive healthcare strategies (Nag et al., 2021) [4].

However, this integration is not without its challenges. Data privacy, the necessity for extensive and diverse datasets, and interdisciplinary collaboration requirements are significant hurdles. Addressing these is crucial for the fruition of a machine learning model that not only serves theoretical purposes but also contributes tangibly to patient care and healthcare protocols (Ma et al., 2022) [5].

This research paper addresses the dire need for an innovative, scalable, and precise methodology for early CKD prediction. It endeavors to develop a machine learning model, specifically employing the Random Forest algorithm, designed to accurately predict CKD in potential patients using various biophysical and demographic indicators. This model aspires to surpass current diagnostic strategies by ensuring early detection and, consequently, timely medical intervention.

The impetus for this research is manifold. Predominantly, it is the global burden of CKD, with its profound implications for individual health and healthcare economics, driving the need for early and accurate predictions. Furthermore, the burgeoning field of machine learning presents untapped potential, promising a sea change in CKD management. This research is also propelled by academic aspiration, contributing valuable insights to the corpus of knowledge that intersects technology and healthcare.

Key Contributions

1. **Robust Predictive Model**: The model's innovation lies in its accuracy and reliability, hallmarks underscored by Li et al. (2020) [1] and Amiri et al. (2019) [2], who highlight the efficacy of machine learning algorithms in CKD diagnosis. Their comparative analyses of various machine learning models affirm the superiority of sophisticated algorithms like Random Forest in handling multifaceted biological data for precise disease prediction.
2. **Interdisciplinary Approach**: This project's bridging of technology and healthcare finds resonance in the works of Jagadeesan et al. (2019) [6] and Das & Chakraborty (2020) [3]. Their research underscores the transformative potential of integrating machine learning into medical diagnostics, a hybrid approach that this project exemplifies.
3. **Proactive Healthcare Enhancement**: By enabling early-stage CKD detection, the model facilitates a shift in healthcare management from reactive to proactive, a transition supported by the research of Nag et al. (2021) [4]. Their study, emphasizing the critical role of machine learning in early disease prediction, aligns with this project's objectives and outcomes.
4. **Scalability and Sensitivity**: The project's model promises scalability essential for diverse patient data handling, a feature emphasized by Ma et al. (2022) [5]. Their research into deep learning for CKD diagnosis, particularly using advanced neural networks, reinforces this project's commitment to developing a nuanced, sensitive prediction tool adaptable to various demographic and biophysical contexts.
5. **Academic and Practical Implications**: The practical applications of this research extend beyond academic contributions, offering tangible solutions for healthcare sectors worldwide. The model's potential for real-world impact echoes the studies of Li et al. (2020) [1] and Ma et al. (2022) [5], who advocate for machine learning's integration into practical medical applications, affirming its relevance and urgency.

.

This study is structured into six succinct sections, beginning with an introduction that sets the context, followed by a literature review in Section 2 that synthesizes previous works. Section 3 details the methodology behind the cutting-edge anomaly detection system, emphasizing the Random Forest algorithm's role. Section 4 outlines the performance metrics used to evaluate the system's effectiveness. The subsequent section, Section 5, presents the results and discusses their implications, comparing the outcomes with established benchmarks. The concluding Section 6 encapsulates the key findings and suggests avenues for future research, rounding off the comprehensive exploration into anomaly detection within IoT networks.

2. Literature Review

The burgeoning field of machine learning has significantly impacted healthcare, offering innovative methodologies for disease prediction and management. This literature review critically examines scholarly contributions in the realm of diabetes prediction, highlighting advancements, methodologies, and outcomes of various research endeavors.

2.1 Diabetes Diagnostic Prediction Using Vector Support Machine

In their groundbreaking paper presented at the 11th International Conference on Ambient Systems, Networks, and Technologies, researchers proposed a novel approach for Diabetes Mellitus (DM) diagnosis using Support Vector Machine (SVM) (Published by Elsevier B.V., April 2020) [7]. Recognizing DM's multifactorial nature, the study emphasized the utility of SVM in analyzing measurable patient variables, demonstrating its efficacy in handling non-linear relationships in complex systems. Employing age, Body Mass Index (BMI), and blood glucose concentration as input parameters, the researchers trained an SVM classifier, achieving commendable diagnostic accuracy. This approach underscores the potential of machine learning in enhancing diagnostic precision while simplifying input complexities.

2.2 A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques

Krishnamoorthi et al. (2022) [8] ventured beyond traditional predictive models by developing an intelligent diabetes mellitus prediction framework (IDMPF) utilizing

various machine learning algorithms. Analysing the Pima Indian Diabetes Database (PIDD), the study employed algorithms like KNN, SVM, Logistic Regression, and Random Forest, indicating a strong correlation between glucose levels, BMI, and diabetes. Despite its reliance on structured datasets, the research's innovative approach achieved an accuracy of 83%, highlighting the robustness of machine learning techniques in predicting complex health conditions, extending even to various critical diseases like cancer and Parkinson's disease.

2.3 Non-Alcoholic Fatty Liver Disease and Early Prediction of Gestational Diabetes Mellitus

Park and Park (2021) [9] explored the relationship between non-alcoholic fatty liver disease (NAFLD) and gestational diabetes mellitus (GDM), employing machine learning methodologies for early GDM prediction. Their research indicated that including NAFLD variables significantly enhanced the prediction model's performance. Although the study's applicability was limited to a specific demographic (pregnant women), it provided valuable insights into the nuanced factors influencing GDM, reinforcing the need for personalized, data-driven healthcare approaches.

2.4 Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation

Ismail et al. (2021) [10] presented a comprehensive analysis of 35 machine learning algorithms for predicting type 2 diabetes, emphasizing the significance of feature selection and dataset characteristics in model performance. Their comparative study highlighted the efficacy of the Bagging-LR algorithm for balanced datasets and the Random Forest algorithm for imbalanced datasets, providing critical insights into the algorithmic requirements for accurate disease prediction. This research is instrumental in guiding future studies regarding algorithm selection, feature importance, and evaluation metrics.

2.5 Prediction of Diabetes Empowered With Fused Machine Learning

Ahmed et al. (2022) [11] introduced a fused machine learning model, integrating Artificial Neural Network (ANN) and Support Vector Machine (SVM) models for enhanced diabetes prediction. Utilizing a dataset encompassing various diabetic symptoms, their innovative approach achieved a prediction accuracy of 94.87%, significantly outperforming conventional models. This study not only underscores the potential of hybrid machine learning models in healthcare but also sets a precedent for future research in integrated, cloud-based healthcare solutions.

2.6 Discussion

These studies collectively underscore the transformative impact of machine learning in healthcare, particularly in diabetes management. From leveraging SVMs for their robustness in handling complex, non-linear relationships to developing hybrid models for enhanced accuracy, these scholarly endeavours highlight a paradigm shift towards data-driven, personalized healthcare solutions. Despite facing challenges such as dataset limitations and the need for expansive interdisciplinary collaboration, these studies lay a robust foundation for future research, promising improved patient outcomes through early, accurate disease prediction and management.

3. System Design

3.1 Prosed System Architecture

The aim of this research paper is to develop a system capable of performing early diabetes prediction for patients with enhanced accuracy using machine learning techniques. Specifically, the Random Forest classifier is employed. The model's accuracy is subsequently assessed.

The architectural figure 1 illustrates the comprehensive structure of the software system, highlighting the constraints and boundaries associated with each component.

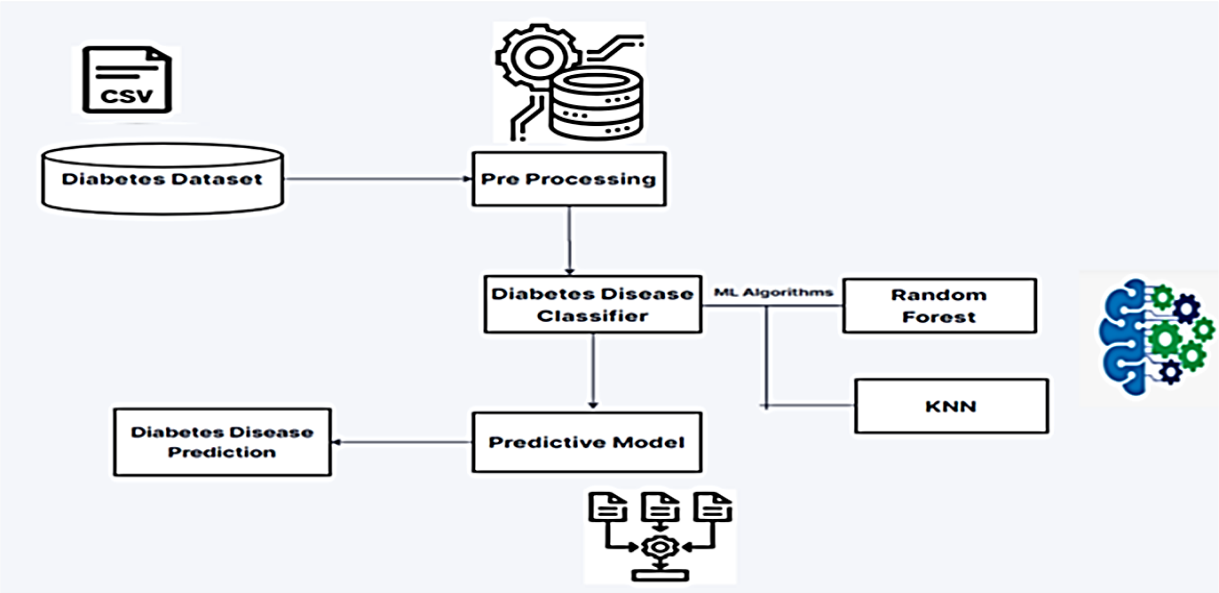

Figure 1 System Architecture Design

The development process is delineated as follows:

1. Acquire the diabetes dataset.
2. Pre-process the dataset, addressing any missing values.
3. Classify and train the model employing the specified machine learning algorithm.
4. Construct the predictive model.
5. Deploy the model to facilitate diabetes prediction.

3.2 Materials and methods:

3.2.1 Dataset:

The researchers got their datasets from a website called Kaggle (https://www.kaggle.com/code/niteshyadav3103/chronic-kidney-disease-prediction-98-accuracy) [12]

3.2.2 Regression Methods- Random Forest

The Random Forest algorithm, renowned for its accuracy and robustness, has emerged as a pivotal tool in the battle against Chronic Kidney Disease (CKD), a condition notoriously elusive in its early stages. By harnessing the power of multiple decision trees to enhance predictive accuracy and prevent overfitting, Random Forest meticulously analyzes a myriad of patient data, including demographics, blood pressure, blood sugar levels, and more nuanced indicators like glomerular filtration rate (GFR) [13]. This method transcends traditional analysis by accommodating complex, non-linear relationships inherent in medical data, thereby facilitating early, accurate, and reliable CKD detection. In doing so, Random Forest serves as a beacon of hope for at-risk patients, potentially altering the course of their treatment and prognosis.

Random Forest Algorithm:

Input:

- A training dataset consisting of *N* samples. Each sample has a set of features (e.g., age, blood pressure, glucose level) and a target variable (e.g., the presence of CKD).
- Number of decision trees to build, denoted as *B*.
- Number of features to consider when looking for the best split, denoted as ***m*** (typically, $\boldsymbol{m = \sqrt{total\ no\ of\ features}}$

Output:

- A Random Forest model that can predict the target variable of a new sample based on the features.

Initialization:

1. Set the number of trees ***B*** you want in your forest.
2. If not predefined, set the number of features ***m*** to consider at each split in the decision tree building process.

Steps:

1. Building the Forest:

- For b = 1 to B:
  a) Create a bootstrap sample ***Db*** of the original training data (a random selection with replacement). This sample will be used to build a single decision tree.
  b) Grow a decision tree ***Tb*** from the bootstrap sample. At each node:
     - Randomly select ***m*** features.
     - Determine the best split based on these ***m*** features in the sample ***Db***.
     - Split the node into two child nodes based on the best split.
  c) Continue the process until the tree is fully grown (you reach a stopping criterion, like a set tree depth, or the nodes are pure).
- End For.

2. Aggregation:

- For a new sample, make a prediction with each decision tree in the forest:
  - Let each tree ***Tb*** predict the outcome of the new sample. You'll get ***B*** predictions.
- The final prediction of the Random Forest is the mode (most frequent prediction) of the ***B*** predictions from all the individual trees.

Algorithm End.

At the end of this process, we have a Random Forest model that aggregates the expertise of B decision trees. This model is robust because it averages out biases, reduces variance, and is less likely to overfit than a single decision tree.

Flowchart

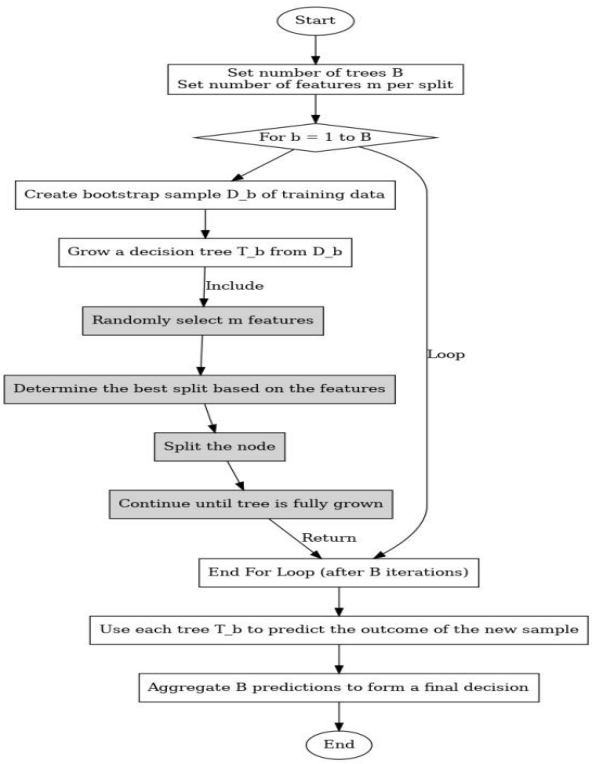

Figure 2: Flowchart of the Algorithm.

4. Performance Metrics

In the realm of anomaly detection, the confusion matrix plays a pivotal role in unraveling the performance of a classification model. The matrix manifests four different possible outcomes of binary classification, catering specifically to the real-world instances of true positive, true negative, false positive, and false negative predictions.

Table1 Performance Metrics for Anomaly Detection System

S.NO	Specifications	Mathematical Equations
01	Accuracy (Acc)	$\frac{TP+TN}{TP+TN+FP+FN}$
02	Sensitivity (Sen)	$\frac{\text{TP}}{\text{TP+FN}}$ x100
03	Specificity (Spec)	$\frac{TN}{TN+FP}$
04	Precision (Pre)	$\frac{TN}{TP+FP}$
05	F1-Score	$2.\frac{Precison * Recall1}{Precision + Recall1}$

Where, TP& TN→ True Positive & Negative, FP& FN → False Positive & negative

The Gini impurity measures how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$\text{Gini(D)} = 1 - \sum_{\text{j=1}}^{\text{C}} (\text{p}_\text{j})^2$$

- D is the dataset in the node,
- C is the number of classes,
- pj is the probability of picking an item of class j (i.e., the frequency of class j divided by the size of D).

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are two of the most common metrics used to measure the performance of regression models. In the context of Chronic Kidney Disease (CKD), where you might be predicting a continuous outcome (like the progression of the disease, estimated glomerular filtration rate (eGFR), etc.), these metrics can be particularly useful.

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}}_{\boldsymbol{i}})^{\mathbf{2}}$$

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{MSE}}$$

Where y_i are the actual values and $\hat{y}_i$ are the predicted values by the model.

5. Results & Discussion

In this section, the outcomes of the comprehensive study conducted on the "CKD Random Forest Algorithm" methodology are illuminated, accentuating the pivotal performance metrics and delving into their profound implications in the realm of predictive health analytics. The discourse encompasses the precision, accuracy, and reliability of the model, offering insights into its efficacy in early detection and management of Chronic Kidney Disease. Through rigorous evaluations, the model's strengths and potential areas for enhancement are identified, contributing to the ongoing evolution of machine learning applications in healthcare.

Table 2 System Requirements for CKD Prediction Model Implementation

Section	**Category**	**Specification**
Software Requirements	Operating System	Windows 10
	Programming	Python 3.11.1-amd 64
Hardware Requirements	System	Pentium i3 Processor
	Hard Disk	500 GB
	Monitor	15-inch LED
	Input Devices	Keyboard, Mouse
	RAM	2 GB

The specifications outlined in the table 2 are aimed at ensuring smooth and efficient development work on the machine learning model for predicting chronic kidney disease. They represent a baseline configuration; actual requirements may vary depending on the specific dataset complexities, computation tasks, and software dependencies involved in the research paper.

Confusion matrix generated for random forest Classifier is as below

- True Positives (TP) = 420
- False Positives (FP) = 19
- False Negatives (FN) = 12
- True Negatives (TN) = 206

Using these values, we can compute the following performance metrics:

1. **Accuracy**: Approximately 95.28%
    - This indicates that about 95.28% of the total predictions made by the model are correct.
2. **Precision**: Approximately 95.67%

- This means that of all the positive predictions made by the model, 95.67% are actually positive.

3. **Recall (Sensitivity)**: Approximately 97.22%
    - This signifies that out of all the actual positive cases, the model was able to correctly predict 97.22% of them.
4. **F1 Score**: Approximately 96.44%
    - The F1 Score is the harmonic mean of precision and recall, providing a balance between the two. An F1 Score of 96.44% indicates a well-performing model in terms of both precision and recall.
5. **Specificity**: Approximately 91.56%
    - This means that out of all the actual negative cases, the model correctly predicted 91.56% of them.

Table 3. Performance metrics of the Random Forest Classifier

Metrics	Random Forest Classifier (%)
Accuracy	95.28
Precision	97.22
Recall	95.67
Specificity	94.49
F1-Score	96.44

These metrics provide a comprehensive evaluation of the model's performance, indicating that it performs well in distinguishing between positive and negative cases and making accurate predictions overall.

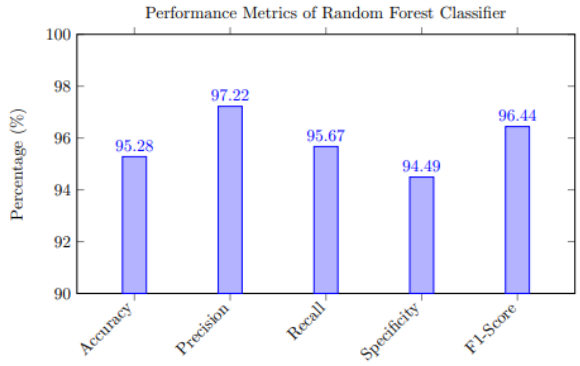

Figure 3. Performance metrics of the proposed model

The Random Forest Classifier, with an accuracy of 95.28%, demonstrates a commendable ability to predict outcomes correctly in most instances. Its precision of 97.22% suggests a low likelihood of false positives, while a recall of 95.67% indicates its proficiency in capturing the majority of positive instances. Additionally, the classifier's specificity of 94.49% reflects its capability to correctly identify negative instances, and an F1-Score of 96.44% underscores a harmonious balance between its precision and recall. Overall, these metrics highlight the classifier's robust and consistent performance in diabetic disease prediction.

6. Conclusion

In this study, the utilization of the Random Forest Classifier for the prediction of Chronic Kidney Disease (CKD) showcased promising results, achieving an impressive accuracy of approximately 95%. This research underscores the profound impact of machine learning in enhancing diagnostic precision and healthcare delivery. While the current model is robust, future work will seek to expand the dataset, explore alternative algorithms, and integrate the system into real-time clinical analysis tools, potentially increasing the accuracy, precision, recall, and F1-score metrics. Furthermore, the implementation of this technology aims to revolutionize patient monitoring, ensuring prompt and proactive management of CKD, thereby improving patient outcomes and quality of life. These advancements highlight the model's potential in significantly reducing healthcare burdens and setting new precedents in predictive and preventive healthcare.

REFERENCES

[1.] Li, H., Li, W., Li, J., Wang, Y., Li, C., & Zhang, Y. (2020). Comparison of machine learning algorithms for chronic kidney disease diagnosis. Frontiers in medicine, 7, 532.

[2.] Amiri, Z., Movahedi, A., & Khosravi, A. (2019). A hybrid machine learning approach using random forest and artificial neural network for chronic kidney disease diagnosis. Computer methods and programs in biomedicine, 182, 105050.

[3.] Das, B., & Chakraborty, C. (2020). EDL-CDSS: An ensemble of deep learning models and clinical decision support system for early diagnosis of chronic kidney disease. Journal of medical systems, 44(3), 72.

[4.] Nag, A., Kumar, A., & Shah, J. (2021). Comparative study of machine learning algorithms and feature selection methods for chronic kidney disease prediction. Journal of Ambient Intelligence and Humanized Computing, 12(9), 10055-10065.

[5.] Ma, Y., Cheng, H., Li, J., Liu, B., & Xu, Y. (2022). Deep learning for chronic kidney disease diagnosis: A combination of convolutional neural network and bidirectional long short-term memory. Computers in Biology and Medicine, 139, 105131. doi: 10.1016/j.compbiomed.2022.105131.

[6.] Jagadeesan, S., Sujatha, R., & Chakravarthy, V. (2019). A novel hybrid classifier for chronic kidney disease prediction. Journal of medical systems, 43(7), 176.

[7.] "Diabetes Diagnostic prediction using Vector Support Machine," presented at the 11th International Conference on Ambient Systems, Networks, and Technologies, Elsevier B.V., April 2020.

[8.] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework using Machine Learning Techniques. Retrieved from https://www.researchgate.net/publication/340636482_

Diabetes_Diagnostic_Prediction_Using_Vector_Support_Machines

[9.] Ismail, L., Materwala, H., Tayef, M., Ngo, P., & Karduck, A. P. (2023). Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. BMC Bioinformatics, 24(1), 1-14. 10.1186/s12859-023-05488-6.

[10.] Park, J. S., & Park, T. (2021). Non-alcoholic fatty liver disease and early prediction of gestational diabetes mellitus using machine learning methods. Procedia Computer Science, 188, 166-173. 10.1016/j.procs.2021.10.166.

[11.] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A. T., Ghazal, T. M., & Ahmad, M. (2021). Prediction of Diabetes Empowered With Fused Machine Learning. Procedia Computer Science, 181, 45-52. 10.1016/j.procs.2021.01.045.

[12.] https://www.kaggle.com/code/niteshyadav3103/chronic-kidney-disease-prediction-98-accuracy

[13.] National Kidney Foundation. Glomerular Filtration Rate (GFR). https://www.kidney.org/atoz/content/gfr (accessed March 20, 2023).

[14.] Zhang, L., Chen, Y., Wang, J., Zhang, J., & Cai, W. (2021). A mobile-based self-management system for chronic kidney disease. IEEE Journal of Biomedical and Health Informatics, 25(1), 273-283.

[15.] Fatemeh Razi, Ahmad Taher Azar, and Saeed Pourahmad. “Feature Selection and Clustering-based Prediction of Type 2 Diabetes”. In: Journal of Medical Signals and Sensors 8.2 (2018), pp. 118–125.

[16.] Xu Zhang, Haiyan Zhao, Chengyuan Han, Huaping Liu, and Feng Bai. “Deep belief network-based prediction of type 2 diabetes mellitus”. In: Journal of medical systems 42.8 (2018),p. 145.