Prediction of Renal Illness using Machine Learning Models

CHANDRA SEKHAR SANABOINA [1]
SRI SATYA PREM CHARAN KURELLA [2]

University College Of Engineering, Kakinada
Department of Computer Science And Engineering
Kakinada - Andhra Pradesh - India
[1]chandrasekhar.s@jntucek.ac.in
[2]charankurella1999@gmail.com

Abstract. The biggest issue that the entire globe may encounter at this time is chronic kidney disease. Early stages are symptomless and only become apparent when kidney function has been reduced by up to 25%. Therefore, it is necessary to anticipate and detect chronic renal disease. Due to their rapid and precise detection capabilities, machine learning models are employed nearly exclusively in clinical and medical settings to identify a variety of chronic conditions. Here Chronic kidney Disease dataset is used from the UCI repository, several machine-learning algorithms are used in order to predict various chronic diseases. The proposed system uses a Stochastic Gradient Descent algorithm to make our model learn a lot faster. The expected results will be a comparative table for various machine learning algorithms with respect to performance metrics like Precision, F1-score, Recall, and Accuracy.

Keywords: Chronic kidney disease, Stochastic Gradient Descent, Feed Forward Neural Networks, Machine Learning, Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine, K Nearest Neighbour.

(Received May 12th, 2022 / Accepted June 1st, 2023)

1 Introduction

A successful existence is closely related to being in excellent health. The body's numerous organs interact with one another to function. The organs must be in good health in order to perform at their highest level. Being in great health is important since it pertains to the condition of being socially, psychologically, and physically well. Hence it is necessary to take care of every organ in our body to lead a healthy life. In contrast to many other diseases, chronic kidney disease CKD may take a long time for its consequences to become apparent in a patient. CKD is an asymptomatic disease at its earlier stage, with few or no symptoms, sometimes without disease-specific symptoms, making it difficult to predict, recognize and prevent. The sooner a disease is detected, the sooner treatment can begin, which, if undetected, can lead to permanent health problems and even death. Machine learning is expected to provide a low-cost detection and prediction solution for this medical issue (12; 11). This may help doctors diagnose more patients more rapidly by allowing them to begin handling CKD patients at an early stage. symptomless individuals are screened for CKD to enable earlier therapeutic intervention and prevent improper exposure to nephrotoxic substances, both of which have the potential to considerably reduce the course of CKD to end-stage renal disease. Machine learning can predict the occurrence, course, and determinants of individual chronic diseases in many contexts (13; 2). The results are unique and relevant to improving clinical decision-making and the organization of healthcare facilities. Machine learning accelerates data processing and analysis (10). With machine learning, predic-

tive analytics algorithms can be trained on even larger datasets and easily modified at deployment time to perform deeper analysis and prediction of various chronic diseases. This study employs machine learning (ML) methods to categorize and predict CKD. Healthcare facilities, stakeholders, and experts will find it simpler to identify and categorize patients as having CKD or not as a result.

2 Literature Review

The kidneys will eliminate extra water and waste from the blood. As the kidneys deteriorate waste builds up which results in the development of symptomless disease. Laboratory testing can still diagnose a patient even if they have no symptoms at all. Symptoms can be treated with medications. Later stages may call for mechanical hemofiltration dialysis or transplantation. A. A. Johari et al. (7) used two classification algorithmsâtwo-class decision trees two-class neural network algorithmsâto compare two algorithms for the anticipation of chronic renal disease. Out of these, the neural network outperformed the decision tree in terms of accuracy 99.56%. J. D. De Guia et al. (8) deal to anticipate chronic kidney disease, the following machine learning classifiers were employed in the study: Random Forest, ANN, Naive Bayes, Decision Trees, SVM, and MLP. Out of the six algorithms, the ANN algorithm has the highest F1 score of 0.992248 and the quickest training time of 46.999 ms. R. Gupta et al.(5) performed a performance analysis on a number of machine-learning methods for chronic renal disease prediction. Decision trees, logistic regression, and random forest are the algorithms. Among these, logistic regression, decision tree, and random forest all attained accuracy levels of 98.48, 94.16, and 99.24, respectively. I. U. Ekanayake et al.(1) Used 11 machine learning classifiers to detect chronic kidney disease in its early stages. RF, XGBoost, LR, SVM, AdaBoost, KNN, NEURAL NETWORK, GNB, DECISION TREE to support the early detection and treatment of patients in order to save their lives. Random forest outperformed the others, with a 99.85% accuracy rate. B. Gudeti et al.(4) made a research on the different algorithms and compare them using different performance criteria. SVM, KNN, and Logistic Regression were the models used in their research. The Support Vector Machine outperformed them all with an accuracy of 99.25%. The advantage of this strategy is that the prediction procedure requires far less time, allowing clinicians to start treating CKD patients as soon as possible. S. Y. Yashfi et al.,(15) proposed a technique for estimating the probability of CKD. 455 patientsâ data were used in this study. It utilizes both an online dataset provided by the Khulna City Medical College's real-time dataset as well as the UCI Machine Learning Repository. After the data was trained with a 10-fold CV, random forest and ANN were employed in this instance. The accuracy of the random forest approach is 97.12%, whereas the accuracy of the ANN is 94.5%. This strategy helps with the early diagnosis of chronic renal disease prediction. P. Ghosh et al.(3) deals with handling the entire study and providing extremely accurate prediction results of CKD, Here AdaBoost, Gradient Boosting, Linear Discriminant Analysis, and Support Vector Machine have all been used. These algorithms are applied to the online dataset of the UCI machine learning repository. Results from Gradient Boosting GB Classifiers have a predictably high accuracy of about 99.80%. A. Vijayalakshmi et al.(14) used Machine Learning ML classification methods to predict the value in a study to identify the presence or absence of CKD in the patient. Several categorization systems can predict the patient's CKD and non-CKD status. This survey has covered different ML algorithms used to identify renal disease as well as the key problems, which are briefly addressed. The random Forest ML algorithm has shown the best performance, with an order to maintain consistency of 99.75%. M. A. Islam et al.(6) used six algorithms Decision Tree, Random Forest, Simple Logistic Regression, Naive Bayes, Simple Linear Regression Model, and Linear Regression to predict the risk factor for CKD. The Random Forest method yields a high accuracy of 98.8858%, according to an analysis of the findings. G. Nandhini et al.(9) primarily used various machine learning classifiers to provide a successful treatment for early disease prediction. Ensemble classifiers, which combine the anticipated outcomes of various classifiers, help the model perform even better. It used the four-ensemble algorithm, which combines AdaBoost, Gradient Boosting, Random Forest, and bagging. The effectiveness of these classifiers was measured using a variety of metrics. In terms of accuracy, AdaBoost and Random Forest did better with 100% Accuracy.

3 Preliminaries

This section describes the preparations before building the model, including a description of the dataset, Operating environment, and metrics used for comparing the performance of various models

3.1 Details of the data and the working environment

The hospital data collection by Soundarapandian et

al. in the UCI machine learning repository provided the CKD dataset for this study. 400 samples make up the dataset. The 24 predictors or features are classified as 11 categorical and 13 Numeric values. The Dataset consists of attributes like sugar, blood pressure, etc. There are two classes for the output variable i.e., CKD for +ve symptoms and not CKD for -ve symptoms. Out of 400 samples, 250 were classified as having CKD and 150 as not having CKD. There are a few missing values in the data. The missing values of attributes are to be filled for better analysis.

3.2 Data processing

For ease, each nominal (categorical) variable is coded and processed by a computer. Medical terms like PCC, ba, htn, dm, etc are in a categorical format which are encoded as 0 or 1. Even though the three variables sg, al, and sc are categorical variables by definition, their values are still numbers. Therefore, these variables were treated as numbers.

The samples range from 1 to 400. When patients did not consult the diagnostic center for which the dataset may miss some medical diagnosis values. Since the number of samples is uncertain, an appropriate imputation is required. The original CKD datasets missing values were handled and filled once the categorical variables were encoded. For filling in the missing values KNN imputation was used which works on the principle of choosing the nearest K samples and selecting the one with the smallest Euclidean distance. Samples of numeric type are filled with the median

3.3 Performance Measure

CKD and non-CKD were chosen to be positive and negative in this study, respectively. A true positive (TP) indicates that the diagnosis of the CKD sample was correct. False negatives (FN) indicate that CKD was misdiagnosed in samples. A false positive (FP) indicates that the model failed to identify CKD. True negative (TN) indicates that the Notckd probe's diagnosis was correct. Precision, Accuracy F1Score, and recall were used to assess the model's performance. These are calculated using the following formulas:

3.3.1 Accuracy

Accuracy is one of the most straightforward metrics to evaluate and is determined as the ratio of accurate predictions to all other guesses. One way to put it is as shown in eq 1.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

3.3.2 Precision

The proportion of accurately predicted positive observations to all anticipated positive observations is determined by the precision score. It is shown in eq 2.

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

3.3.3 Recall

Recall/sensitivity is a ratio used to compare all observations in a true class to precisely anticipated positive observations. It is shown in eq 3.

$$\text{Recall} = TP/(TP+TN) \quad (3)$$

3.3.4 F1-Score

The weighted average of the Precision and Recall calculations is the F1 score. It can be proven that this score, which accounts for false positive and false negative readings, is more valuable than accuracy. It is shown in eq 4.

$$\text{F1-Score} = (2(\text{Recall}*\text{Precision}))/(\text{Recall}+\text{Precision}) \quad (4)$$

4 Proposed Model

To diagnose the data samples in this section, several machine-learning algorithms were used. The models that performed the best among these were chosen as prospective components. Their errors in judgment were analyzed, and the component models were identified. Next, a stochastic descent gradient model was applied to produce better results.

4.1 Setting up and evaluating initial individual models

On the entire CKD data sets, the corresponding subset of features or predictions are applied, and the following machine learning models were used with the goal of identifying CKD

1) Logistic regression model
2) Model based on trees: RF
3) SVM, a decision-plane-based model
4) KNN, a distance-based model
5) Model based on probabilities: NB

6) Feed Forward Neural Network

4.2 Establishing the stochastic gradient descent

To identify model parameters that offer the best fit between expected and actual outputs, machine learning applications frequently use the stochastic gradient descent optimization process. This approach works but is unreliable. Stochastic gradient descent is widely used in the machine learning sector. As a result, stochastic gradient descent chooses a subset of the dataset at random for each iteration. The "batch" in gradient descent refers to the number of samples from the dataset used to compute the gradient for each iteration. In a typical gradient descent optimization like batch gradient descent, the batch is viewed as the entire dataset. Making use of the entire dataset can be highly beneficial in reducing noise and unpredictability to a minimum, however, issues occur as the dataset expands. For example, your dataset has 1 million samples. In order to use the gradient descent technique, he must do one iteration for every million samples and continue it until the minimal value is attained. Therefore, the implementation requires a lot of computer resources. Utilizing stochastic gradient descent, this issue is resolved. SGD only uses one probe. A single-stack operation for every iteration. Iterations are carried out by selecting and shuffling samples at random. In fig. 1, the architecture of the proposed system is depicted.

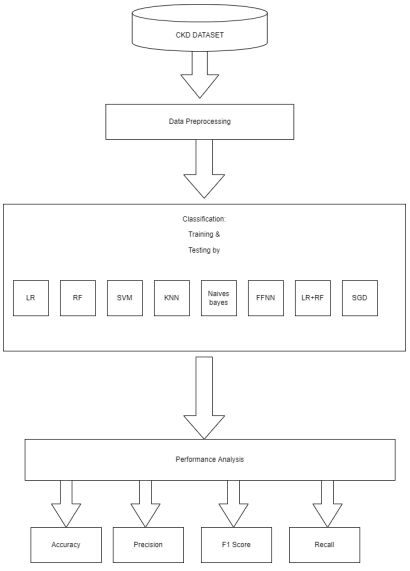

Figure 1: Architecture of Proposed system

4.3 Description of Algorithms

4.3.1 Logistic Regression: In the category of supervised learning techniques, logistic regression is one of the most used machine learning algorithms. It is utilized to forecast a categorical dependent variable using a certain set of independent variables. By using logistic regression, the output of a dependent variable with a categorical component is predicted. The results must be discrete or categorical.

4.3.2 Random Forest: Accurate prediction and improved generalization are made possible by the use of random sampling and ensemble procedures in RF. Many trees make up a random forest. The accuracy increases with the number of decorrelation trees. Some of the missing data can be filled in by a random forest classifier.

4.3.3 Support vector machine: Cortes and Vapnik developed the Support Vector Machine (SVM), a technique for supervised machine learning. The goal of SVM is to determine the best decision boundary with a maximum margin hyperplane between samples of various classes. The SVM must convert the input data space dividing the dataset from a low-dimensional space to a high-dimensional space into multiple samples with optimum boundaries.

4.3.4 K- Nearest Neighbour: Thomas Cover developed the K Nearest Neighbors (KNN) supervised technique to solve classification and regression problems. To predict labels for newly provided points, use the feature similarity technique. Additionally, this implies that new test points are classified according to the agreement of the training set's K nearest neighbors, where K is the number of neighbors.

4.3.5 Feed Forward Neural Network: The first and most basic artificial neural network design was the feedforward neural network. The information in this network only travels in one direction, forward, from the input nodes to the output nodes, passing via any hidden nodes that may exist. The network doesn't contain any loops or cycles.

4.3.6 Naive Bayes: A classification algorithm using Bayesian at its core is known as a naive Bayes classifier. It's a group of algorithms with related definitions, not a single algorithm.

4.3.7 Integrated Model: This model is formed by

the combination of Logistic Regression and Random Forest using ensemble techniques.

4.4 Encoding, Missing values, and Outlier Treatment:

The label encoder will be used to convert the categorical columns' values from categorical to numeric after they have been imputed with KNN imputation. When all of the columns in the complete data frame have been converted to numeric columns, to impute the missing values, we have used the multiple imputations by chained equations (MICE) package. The interquartile range will then be used to find outliers and should be avoided to produce the final working dataset.

4.5 Training and Testing: We split the dataset into training 75% and test 25% sets in order to train and test the CKD prediction model.

4.6 Model Prediction:

Several machine-learning approaches were employed to create a prediction model. The eight techniques we employed includes Random Forest, KNN, Logistic Regression, SVM, FFNN, Naive Bayes, SGD, and Integrated Classifiers.

4.7 Model Comparison:

The model that performed the best in terms of recall, F1-Score, accuracy, and precision must now be chosen.

A comparative table of various ML models over the performance metrics is shown in table (1).

Table 1: Comparative table of various ML models

Model	*Accuracy*	*Recall*	*F1Score*	*Precision*
RF	98.5	99	98	97
KNN	98.5	96	97	94
LR	99.5	99	99	99
NB	95.5	95	94	93
SGD	92.5	83	83	80
FFNN	98.5	98	98	97
ING	99	98	99	97
SVM	99.5	99	99	99

5 Experimental Results

For reaching high accuracy, positive and negative characteristics are portrayed as being more crucial. The most effective classifiers in this research are SVM and Logistic Regression with an accuracy score of 99.5%.

The comparison graphs for various metrics vs various classifiers were shown in Figures (2), (3), (4), and (5).

Figure 2: Accuracy graph of various ML models for CKD

Figure 3: F1 score of various ML models for CKD

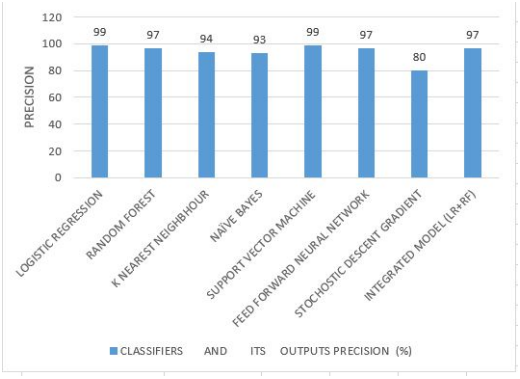

Figure 4: Precision graph of various ML models for CKD

The Integrated Model's achieved the second-highest accuracy result is 99%. The Random Forest, KNN, and Feed Forward Neural Networks classifiers and values have the third-highest accuracy, at 98.5%. The naive

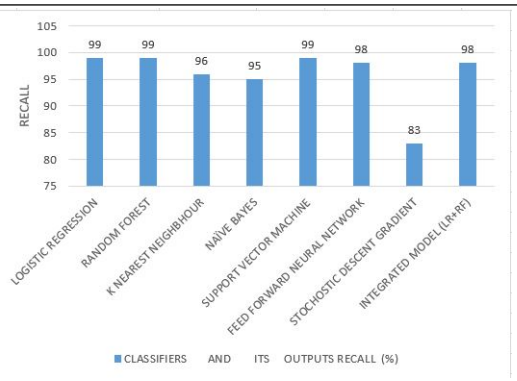

Figure 5: Recall graph of various ML models for CKD

Bayes classifier achieved a value accuracy of 95.5%. We reach 92.5% accuracy with the Stochastic Gradient Descent SGD classifier, which is the second-lowest accuracy. The research's findings section aims to identify each classifier's best attempt.

6 Conclusion

Renal failure is the main cause of death in people with CKD. Chronic renal disease in general is a serious issue for human health. Everyone should be concerned about their health to avoid this at an early stage. We handled the missing data, trained it, and created models for logistic regression and support vector machines. These two algorithms were created in Python. The accuracy we get using the Support Vector Machine and Logistic Regression algorithms are 99.5%, which is a comparatively high level of accuracy.

7 Future Scope

The future direction of renal disease prediction systems using huge amounts of patient data can be accelerated and made more accurate by using machine learning techniques.

References

[1] Ekanayake, I. U. and Herath, D. Chronic kidney disease prediction using machine learning methods. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 260–265. IEEE, 2020.

[2] Ferreira, D. D., Santos, L. O., Alvarenga, T. A., Rodríguez, D. Z., Barbosa, B. H. G., Ferreira, A. C. B. H., dos Santos Alves, D. F., Carmona, E. V., Duran, E. C. M., and de Moraes Lopes, M. H. B. Applications of digital and smart technologies to control sars-cov-2 transmission, rapid diagnosis, and monitoring. In *Omics Approaches and Technologies in COVID-19*, pages 405–425. Elsevier, 2023.

[3] Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., and Khan, A. A. Optimization of prediction method of chronic kidney disease using machine learning algorithm. In *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6. IEEE, 2020.

[4] Gudeti, B., Mishra, S., Malik, S., Fernandez, T. F., Tyagi, A. K., and Kumari, S. A novel approach to predict chronic kidney disease using machine learning algorithms. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1630–1635. IEEE, 2020.

[5] Gupta, R., Koli, N., Mahor, N., and Tejashri, N. Performance analysis of machine learning classifier for predicting chronic kidney disease. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–4. IEEE, 2020.

[6] Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., and Hossain, S. Risk factor prediction of chronic kidney disease based on machine learning algorithms. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 952–957. IEEE, 2020.

[7] Johari, A. A., Abd Wahab, M. H., and Mustapha, A. Two-class classification: Comparative experiments for chronic kidney disease. In *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pages 789–792. IEEE, 2019.

[8] Justin, D., Concepcion, R. S., Bandala, A. A., and Dadios, E. P. Performance comparison of classification algorithms for diagnosing chronic kidney disease. In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–7. IEEE, 2019.

[9] Nandhini, G. and Aravinth, J. Chronic kidney disease prediction using machine learning tech-

niques. In *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pages 227–232. IEEE, 2021.

[10] Okey, O. D., Maidin, S. S., Lopes Rosa, R., Toor, W. T., Carrillo Melgarejo, D., Wuttisittikulkij, L., Saadi, M., and Zegarra Rodríguez, D. Quantum key distribution protocol selector based on machine learning for next-generation networks. *Sustainability*, 14(23):15901, 2022.

[11] Okey, O. D., Melgarejo, D. C., Saadi, M., Rosa, R. L., Kleinschmidt, J. H., and Rodríguez, D. Z. Transfer learning approach to ids on cloud iot devices using optimized cnn. *IEEE Access*, 11:1023–1038, 2023.

[12] Ribeiro, D. A., Melgarejo, D. C., Saadi, M., Rosa, R. L., and Rodríguez, D. Z. A novel deep deterministic policy gradient model applied to intelligent transportation system security problems in 5g and 6g network scenarios. *Physical Communication*, 56:101938, 2023.

[13] Teodoro, A. A., Silva, D. H., Rosa, R. L., Saadi, M., Wuttisittikulkij, L., Mumtaz, R. A., and Rodriguez, D. Z. A skin cancer classification approach using gan and roi-based attention mechanism. *Journal of Signal Processing Systems*, 95(2-3):211–224, 2023.

[14] Vijayalakshmi, A. and Sumalatha, V. Survey on diagnosis of chronic kidney disease usingmachine learning algorithms. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 590–595. IEEE, 2020.

[15] Yashfi, S. Y., Islam, M. A., Sakib, N., Islam, T., Shahbaaz, M., Pantho, S. S., et al. Risk prediction of chronic kidney disease using machine learning algorithms. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2020.