IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12206

A Survey on Chronic Kidney Disease

GANGAMBIKA G[1], MEGHANA K S (1EW19CS075)[2], KAVYA S JAIN (1EW19CS058)[3], POORNIMA M (1EW19CS106)[4], SHWETHA NAIK N (1EW19CS146)[5]

Asst. Prof, Computer Science and Engineering, East West Institute of Technology Bangalore-91[1]

Computer Science and Engineering, East West Institute of Technology Bangalore-91[2-5]

Abstract: People now commonly suffer from chronickidney disease (CKD). By detecting and treating those who are at risk for this condition as soon as feasible, a variety of serious problems, such as end-stage renal disease, elevated risk, and cardiovascular disease, maybe prevented. Medical researchers can get a lot of help from the machine learning algorithm in accurately diagnosing the disease at the very beginning. Algorithmsfor machine learning and Big Data platforms have recently been combined to improve healthcare. This work presents hybrid machine learning methods that integrate extraction of the feature strategies and various algorithms of machine learning under classification technique related to massive data platforms to identify chronic kidney disease (CKD). In this study, logistic regression (LR), random forest (RF), decision tree (DT), support vector machine (SVM), Naive Bayes (NB), and gradient boosted trees were employed as six ensemblelearning strategies for machine learning classificationtasks (GBT Classifier). The results were validated using four evaluation techniques: accuracy, precision, recall, and F1-measure. The results demonstrated that the chosen features had helped SVM, DT, and GBT Classifiers operate at their peak levels.

Keywords: Chronic kidney, Naive Bayes (NB), decision tree (DT), logistic regression (LR), Gradient- Boosted Trees (GBT Classifier) and Random Forest (RF).

I. INTRODUCTION (*HEADING 1*)

In terms of the current state of society's health, this Chronic Kidney Disease (CKD) is viewed as a serious hazard. Regular laboratory testing can identify chronic kidney disease, and there are therapies available that can delay the onset of the condition, stop it from progressing, lessen the risk of cardiovascular disease and its complications, as well as enhance survival and quality of life. Lack of water intake, smoking, a poor diet, insufficientsleep, and many other factors can lead to CKD. In 2016, this condition afflicted 753 million people worldwide, 417 million of them were female and 336 million were male. The majority of the time, the illness is discovered at its advanced form, which can occasionally result in renal failure.

A significant problem for the medical profession is the early detection and treatment of CKD. The treatingphysician (nephrologist) is called upon to treat the aforementioned systemic signs in addition to slowing the disease's progression to more advanced stages and, if possible, suspending it. [7]. ML techniques have become essential tools for a variety of applications in the health sector, including the early identification of some chronic conditions, thanks to the growth of sensor networks, communication technologies, data science, and statistical processing. A machine learning-based strategy for the CKDdisease will be given in the current study endeavour. The following are the primary benefits of the methods used:

- A phase in data pre-processing that makes use of the Synthetic Minority Oversampling Technique (SMOTE), which is necessary to guarantee that the dataset instances are distributed evenly and, as a result, creates efficient classification models to forecast the risk for the development of CKD.
- A features study that consists of three distinct sub-steps: (i) statistical description of numerical attributes, (ii) measurement of relevance using three separate approaches,and (iii) tabulation of frequency of occurrence for nominal features.
- The performance of several models is compared and evaluated using the most used metrics, such as Precision, Recall, F-Measure, and Accuracy. Two kidneys, which are essential organs for the body's healthy operation,are placed in the peritoneal cavity in the rear of the human body. The primary job of the kidneys is to maintain ahealthy equilibrium of water, ions, acids, calcium, phosphorus, magnesium, potassium, and other trace components in the body. The kidneys also release hormones like erythropoietin, vitamin D, and renin at the same time. Erythropoietin primarily promotes the development and maturation of red blood cells in the bone marrow, whereas vitamin D controls the body's levels of calcium and phosphorus, as well as bone structure andmany other processes.

DOI: 10.17148/IJARCCE.2023.12206

Additionally, hormones that control blood pressure, fluid balance, bone metabolism, andvascular calcifications work through the kidneys. Last but not least, the kidneys flush out all waste products of metabolism, medications, and other pollutants that enterthe body. The main conclusion of this investigation can be inferred from a performance evaluation in which all models showed extraordinarily high performance, with Rotation Forest earning the greatest results across all criteria.

The remainder of the work is divided into the following sections. We discuss related works that use ML to exploit the CKD state in Section 2. In addition, we provide the dataset and examine the chosen methodology in Section 3. Section 4 discuss about the how to evaluate the result usingevaluation matrix on different algorithm. Section 5 concludes the essay and lays out the paper's future course.

II. LITERATURE SURVEY

2020 [J. Snegha] [10] suggested a system that makes use ofback propagation neural networks and the Random Forest algorithm, among other data mining approaches. Here, theyevaluate the two algorithms and discover that the Back Propagation model produces greatest results since it makes use of the Feedforward Neural Network, a supervised learning network.

[Mohammed Elhoseny, 2019] claims that a strategy forCKD makes use of ACO and density-based featureselection. Through wrapper methods, the system chooses features.

The creation of a CKD prediction system utilising MLmodel like as Support Vector Machine, Naive Bayes, K- Nearest Neighbor, Logistic Regression, Multi-Layer Perceptron Algorithm, Decision Tree, Random Forest, and was proposed by Baisakhi Chakraborty [9] in 2019. These are employed, and the potency of each is assessed in lightof the precision, accuracy, and recall results. Random Forest is finally used as it gets the better result compared toall other algorithm.

[Arif-Ul-Islam, 2019] proposed a method to predict illness using J48 Decision Tree, Ant-Miner, and Boosting Classifiers. The two objectives in this study, are to evaluateboosting algorithms' efficacy in CKD detection and to create rules that illustrate relationships among the various CKD characteristics. Findings from the trial showed that AdaBoost performed marginally worse than Logit Boost.

For the purpose of predicting CKD, a system utilising an extreme learning machine and ACO has been presented [S.Belina V, 2018]. For classification, a MATLAB tool is utilised, and ELM has certain optimization restrictions. This approach enhances the SLFNs' sigmoid additive type.

2018's SiddheshwarTekale [8] described a machine learning system that makes use of SVM, Decision treealgorithms. By contrasting the 2 methods, it was determined that SVM produces the best results. In order forclinicians to examine patients more quickly, its prediction procedure requires less time.

This paper uses the Back Propagation model of NeuralNetwork technique for prediction was reported in [Nilesh Borisagar, 2017]. Discussed here are Bayesian regularization, Levenberg, Scaled Conjugate, and therobust back propagation algorithm. For the purposes ofimplementation, Matlab R2013a is employed. Scaledmerge gradient and back propagation to get more effective results than Levenberg and Bayesian regularization in terms of training time.

2017 [Guneet Kaur] [7] presented a method for forecasting CKD using Hadoop's data mining algorithms. They employ two KNN and SVM-based data mining classifiers. Here, the manually chosen data columns are used to do the prediction analysis. In this system, SVM classifier provides better accuracy than KNN.

III. PREPARE YOUR PAPER BEFORE STYLING

Figure 1 illustrates the two basic approaches that make up the proposed system for forecasting chronic kidney disease. The first method chooses the most important features from the datasets on chronic renal disease using feature selection methods. The second method uses ensemble learning, DT, RF, SVM, LR, NB, and ML algorithms on this whole set of data and the selected features to predict CKD. Data collection is the first stage in the suggested strategy, which involves six steps and uses the dataset CKD which is from UCI machine learning library.

Null values are handled atthe data preprocessing stage, which is the second step. The third stage will involve choosing the key features usingfeature techniques. The parameters of ML and ensemble learning algorithms are optimized in the fourth stage make use of stratified cross-validation. The following subsections contain a detailed explanation of each stage.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12206

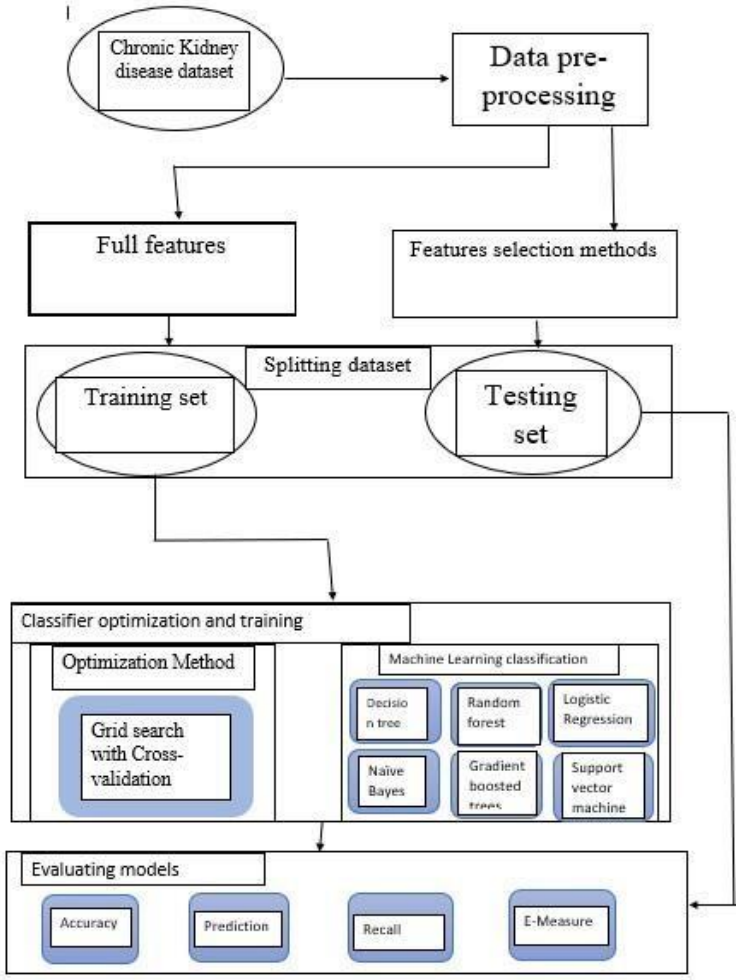

Fig1: Architecture of Chronic Kidney Disease

A. Dataset

The Dataset in CKD from the UCI repository is the dataset we are utilizing in this instance. 400 samples from twoseparate classes are included. There are 25 qualities total, 11of which are numerical, 13 nominal, and 1 class attribute. There are several missing values in the data set. Here, the dataset holds the data about the patient such as blood pressure, age, sugar, RBC count, specific gravity, albumin etc.,

High blood pressure and diabetes can contribute to CKD. Our numerous organs are impacted by diabetes, andexcessive level of diabetes will follow. Therefore, it is crucial to find the illness as soon as feasible. In order to anticipate the sickness, this study improvises a number of machine learning approaches.

B. Pre-Processing

The process of converting distorted or encoded data into a format that can be quickly and easily analysed by a machine is known as data pre-processing. A collection of data elements can be seen as a dataset. The attributes that are considered as an object, like the bunch of a physical object or the precise moment where we can see the event has occurred, are secured by a variety of qualities that are used to identify data items.

Missing values in the dataset may need to be approximated or deleted. Filling in missing values with the mean, median, or mode value of the corresponding feature is the most typical approach of handling missing data. We must change object-typed numerical values to float64 types since object values cannot be used for analysis.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12206

The attribute value currently appearing most frequently in that attribute column is used to replace the null values for the category attributes. By assigning each distinct attribute value to an integer, label encoding converts category data into numeric attributes. This converts the characteristics to an int type automatically. Every missing value in that attribute columnis replaced by the mean value, which is calculated beforehand from each column. For this function, which calculates the mean value for each column, we're utilising afunction called imputer. When the replacement and encoding are finished, The data needs to be evaluated, validated, and trained. The process of actually training our algorithms to create a model occurs during the training of the data phase. The dataset's validation section is used to check the precision of our multiple model fits or to improve the model. Our model's premise is tested using thedata.

C. Feature Selection

The ability to identify the crucial features in a dataset is one of the key advantages of feature selection algorithms. Different classification techniques are used to get betterresults and the model's execution time is decreased with the help of proper feature selection made. The chi-squared feature selection and Relief-F methods accustomed to identify the group of significant characteristics taken from a database. acronyms and abbreviations

D. Splitting the Dataset

For training, 80% of the CKD datasets are used, and for testing, 20%. Utilizing stratified cross-validation, the models were trained and optimized, and the cross-validationoutcomes were documented. Results from the testing set, which we utilized to assess the models, have been recorded.

Models' Optimization and Training

A. Optimization Methods

The models are improved and the hyperparameters are tunedby using grid search with K-Fold cross-validation. The usually used technique for tuning parameter optimization is grid search. Each tuning parameter should first have a collection of principles established by users. Then model makes a choice of the hyperparameter with the highest performance after evaluating each possible value for the hyperparameters.

The dataset is partitioned into k folds of equal size for K- Fold cross validation. Hence, the classifiers are tested in the remaining time after the k-1 group training has been completed. This procedure is performed after the 10 folds have each been presented used as a test set. The efficacy of the classifiers for each k is also assessed. Finally, the evaluation classifier is developed based on average performance.

Machine Learning Models

Here is a list of the categorization models that were utilized in the study:

• Decision tree (DT): For classification problems including a well-liked target variable, it could be a supervised learning strategy. Each distinct and continuous input andoutput variable is supported by a decision tree. Thisapproach applies decision trees to any classification and regression problem when the sample or population is split into two or more groups that are exactly the same, called subpopulations, to support the main splitter of the input parameter.

• Random forest (RF): It is a technique for supervised computer learning. Simply defined, it gathers a lot of trees and combines them for more precise prediction [23]. Problems with binary categorization were tackled using logistic regression (LR). In order to predict the chance of different labels for an unlabeled observation, LR employs a logistic or sigmoid function [35].

• A supervised machine learning (ML) technique is the support vector machine (SVM). The hyperplane is used to sort the dataset. [22].

• Naive Bayes (NB): Using the Bayes theorem, the NaveBayes approach trains a classifier. In other words, the Naive Bayes method was used to train a probabilistic classifier. It creates a probability distribution for a certain observation over a number of classes. [29].

Gradient-Boosted Trees (GBTs): Another method for training a collection of decision trees is to utilize the algorithm known as Gradient-Boosted Trees (GBTs). Every decision tree, however, undergoes sequential training. In order to optimize each new tree, this method uses data from previously trained trees. Consequently, with each new tree, the model

DOI: 10.17148/IJARCCE.2023.12206

gets better. The training of a model could take longer with GBT since it trains one tree at a time. When a lot of trees are used in an ensemble, overfitting is also a possibility. However, a GBT ensemble may have shallow trees, which facilitates training. Gradient boosting is a method for continual training of many decision trees. This model predicts the label for unit training sample using the current ensemble and compares the forecast to the actual label for each iteration.

IV. EVALUATION MATRIX

Equations 1-4 list four standard metrics that are used to evaluate the models: True positive, true negative, false positive, and false negative are all abbreviated as TP, TN, FP, and FN, respectively. Other indicators include accuracy,precision, recall, and F1-score.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}.$$

V. CONCLUSION

This survey helps to propose a model that helps in chronic kidney disease prediction. Early detection and treatment of CKD can be implemented at low cost, thereby reducing the burden, improving diabetic and cardiovascular disease (including hypertension) outcome and significantly lowering patient morbidity and mortality. Our intention is toprovide an effective system in the simplest way which helpsboth doctors and patients to predict chronic kidney diseaseat early stages. Physicians and radiologists can use a computer-assisted diagnostic system to help them make better diagnostic conclusions. Our method enables doctors to serve a greater number of patients in less time.

Proper feature selection methods aid in reducing the amount of features required by the prediction algorithm, hence reducing the number of medical tests required. For better CKD prediction, Future research should investigate different supervised and unsupervised machine learning techniques, as well as feature selection techniques with additional performance metrics. To collect the most recent data for CKD diagnosis from various regions around the world. The sample size (400 instances) is expected to be small, which may affect the reliability of the studies. As a result, the dataset size must be increased in the future for better accuracy.

REFERENCES

[1] Hussein abbass, “classification rule discoverywith ant colony optimization”, research gate article, 2004
[2] Mohammed deriche, “feature selection using antcolony optimization”, international multi- conference on systems, signals and devices, 2009
[3] X. Yu and t. Zhang, ”convergence and runtime ofan anticolony optimization”, information technology journal 8(3) issn 1812- 5638, 2009
[4] David martens, manu de backer, raf haesen, “classification with ant colony optimization”,ieee transactions on evolutionary computation,vol.11, no.5, 2010.
[5] Vivekanand jha, ”chronic kidney disease globaldimension and perspectives”, lancet, national library of medicine, 2013
[6] Kai-cheng hu, “multiple pheromone table basedon ant colony optimization for clustering”, hindawi, research article, 2015.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12206

[7] Guneet kaur, "predict chronic kidney disease using data mining in hadoop, international conference on inventive computing and informatics, 2017.

[8] Siddeshwartekale, "prediction of chronic kidney disease using machine learning, international journal of advanced research in computer and communication engineering, 2018.

[9] Baisakhi chakraborty, "development of chronic kidney disease prediction using machine learning", international conference on intelligent data communication technologies, 2019.

[10] J. Snegha, "chronic kidney disease prediction using data mining", international conference on emerging trends, 2020.