'National Conference on Advancements in Computer Science and Engineering'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Predicting Chronic Kidney Disease Using Machine Learning Algorithms : A Comparative Study

Vennala M Reddy
PG Student, Department of Computer Science and Engineering, New Horizon College of Engineering, India

ABSTRACT
One of the most fatal diseases that gradually harms human kidneys is chronic kidney disease (CKD). The disease is commonly misdiagnosed in its early stages, and many patients become aware of its gravity only after it has progressed. Therefore, diagnosing such diseases sooner is a major challenge nowadays. One of the newest technologies employed in the healthcare industry for disease diagnostics is machine learning. In order to identify the best machine learning classification strategy for the prediction of CKD, we compute, analyze, and compare various approaches in this study. Several well-known machine learning techniques, including GaussianNB, Logistic Regression, K-Nearest Neighbour Classifier, Decision Tree Classifier, and Artificial Neural Network (ANN), were chosen to train the model.

Keywords: Machine Learning, GaussianNB, Logistic Regression, K-Nearest Neighbour Classifier, Decision Tree Classifier, Artificial Neural Network.

I. INTRODUCTION

The kidneys, two bean-shaped organs, are crucial components of the human body. Filtering by the kidney eliminates waste from the blood. Protein can leak into urine and waste materials can stay in blood if this filtration system is compromised. Eventually, the kidney's filtering capacity is lost. Chronic Kidney Disease (CKD), also known as Chronic Renal Disease, is the term used to describe this kidney failure. Kidney failure affects the entire body. In general, this disease worsens with age, but recently, children and youth as young as 5 years old are also affected. Muscle cramps, nausea, vomiting, loss of appetite, ankle and foot edema, excessive or insufficient urination, and difficulty catching your breath are some signs that your kidneys are failing. According to data from over the last fifteen years, there has been an increase in the number of CKD patients, and more than 60% of these individuals are not receiving medical treatment. The leading cause of death worldwide in 1990 and 2010 was CKD, with positions 27 and 18, respectively. 2013 saw 956,000 deaths as a result of CKD. The patient must undergo kidney transplantation or dialysis at the end of the process. Early therapy is one of the best methods to lower this death rate. However, in developing nations, people only seek care when they are in a bad condition. To find CKD sufferers before they reach the end stage, an automated method can be created. Our goal is to calculate, examine, and contrast machine learning classification methods. Finding CKD at an earlier stage is currently a major difficulty because it is not recognized in the early stages and patients only become aware of the severity of the disease once it has progressed. This initiative promotes

early detection and awareness among the public. The progression of this chronic condition may be stopped or slowed down with early detection and appropriate therapy.

In this comparative study proposed for the missing values, we are employing mean, mode, and median-based per-processing approaches. In addition, we trained the model using K-Nearest Neighbour Classifier, Decision Tree Classifier, Gaussian Naive Bayes, Logical Regression, and Artificial Neural Network. Following that, we may compare and decide which of the following methods can most accurately forecast the possibility of CKD based on the outcomes of each of these machine learning methods.

II. LITERATURE SURVEY

Academics and clinicians are increasingly interested in developing tools and procedures for monitoring and forecasting various diseases, with a concentration on those that occur often in human life. In this section, we'll look at recent studies that use machine learning algorithms to predict CKD risk, as well as strategies for processing small datasets. Existing literature was researched to gain the necessary knowledge about various ideas linked to the current application. The following is a list of some of the noteworthy conclusions that were drawn from those.

In this study, the Random Forest Algorithm, a machine learning technique, is used to predict CKD. A combination of tree predictors called a "random forest" makes all the trees rely on independently sampled random vectors. Although many other algorithms, such as SVM and Naive Bayes, were employed, Random Forest proved to be the most accurate [1]. In this research effort, several machine learning techniques are used to predict CKD. Techniques like Naive Bayes, SVM, MBPN, LDA, and KNN are among them. This study verified that all these strategies have a respectable margin of error for effective prediction [2]. To begin, in [3], the authors' research was based on clinical and blood biochemical measurements from 551 patients with proteinuria, and several predictive models were compared, including random forest (RF), extreme gradient boosting (XGBoost), logistic regression (LR), elastic net (ElasNet), lasso and ridge regression, k- nearest neighbour (k-NN), support vector machine (SVM), and artificial neural network (ANN). The authors of [4] created exceptionally accurate CKD prediction models using SVM, AdaBoost, linear discriminant analysis (LDA), and gradient boosting (GBoost) approaches. These models' performance was evaluated using a dataset derived from the UCI machine learning library. With a score of 99.80%, the gradient boosting classifier was the most accurate. The authors of [5] focused on the [6] dataset, using the LR, Decision Tree (DT), and k-NN algorithms to train three different models for CKD prediction. The LR achieved higher accuracy (97%) than the DT (96.25%) and k-NN (71.25%), respectively.

The researchers of [6] concentrated on the dataset from [28]. Three alternative models for CKD prediction were trained using the LR, Decision Tree (DT), and k-NN algorithms. In comparison to the DT (96.25%) and k-NN (71.25%), the LR achieved higher accuracy (97%). The [7] dataset is also used in [8] research. The authors investigated the performance of the Nave Bayes (NB), RF, and LR models for predicting CKD risk, which attained accuracy of 93.9%, 98.88%, and 94.76%, respectively. Furthermore, in [9], the researchers designed a method for predicting CKD risk using 455 patients' data from the UCI Machine Learning Repository and a real-time dataset from Khulna City Medical College. RF and ANN were trained and tested on the data using 10-fold cross-validation. The RF and ANN have accuracy of 97.12% and 94.5%, respectively.

Furthermore, [10] used a CKD dataset from the UCI repository to train and test multiple classifiers, including ANN, C5.0, chi-square automatic interaction detector, LR, linear SVM with penalty L1 and L2, and random tree (RT). Under SMOTE and all features as input to the ML models, the linear SVM with penalty L2 attained the greatest accuracy of 98.86%. [11] conducted experiments on the CKD dataset, which comprised 25 attributes and was available from the UCI Machine Learning Repository. For the diagnosis of CKD, three ML models were chosen: RF, DT, and SVM, with prediction accuracy of 99.16%, 94.16%, and 98.3%, respectively. Furthermore, the authors looked at a dataset of 26 CKD-related factors in [12]. Four feature-based methods were used with the ANN classifier: Extra Tree, Pearson correlation, the lasso model, and chi-square. The ANN ensemble using the lasso model performed the best (99.98%). Furthermore, in the research effort in [13], the extra-trees (ExTrees) classifier, AdaBoost, k-NN, GBoost and XGBoost, DT, gaussian Nave Bayes (NB), and RF were applied. The best classifiers, according to the results, were k-NN and ExTrees.

III. MATERIALS AND METHODS

A. Overview of machine learning

Machine learning is an application of artificial intelligence (AI) that enables systems to automatically learn from experience and change without the need for custom coding. To find patterns in data and improve future decisions based on the examples we provide, the learning process starts with observations or data, such as examples, firsthand experience, or instruction. The primary goal of machine learning is to enable computers to automatically learn and adapt their behaviour in order to increase the program's accuracy and usefulness without any human support or intervention. The process of building computer programmes traditionally involves automating the actions that must be taken on input data in order to produce output artifacts.

B. Supervised and Unsupervised Learning

The following sorts of machine learning approaches can be generally categorized training set is the collection of feature/label pairings used in supervised learning. The system builds a program with a set of rules that represents a generalized model of the relationship between the set of descriptive features and the target features from this training set. The goal is to utilise the generated output program to predict the label for an unknown, unlabeled input collection of features, or to forecast the result for one new piece of data. The produced model classifies data with known labels that are not part of the training set, and the classification results are compared to the known labels. The test set refers to this dataset. The percentage of accurate predictions the model labelled relative to the total number of events in the test set can then be used to determine the predictive model's accuracy. A dataset of descriptive characteristics without labels is used as the training set in unsupervised learning. Unsupervised learning allows the algorithms to find intriguing data structures on their own. The objective at this point is to develop a model that unearths any hidden structure in the dataset, such as relationships or clusters that occur naturally. Unsupervised learning investigates how systems might extrapolate a function from unlabeled data to describe a hidden structure. Although the system is unable to determine the proper output, it explores the data and can infer hidden structures from unlabeled data using datasets. Clustering, which is used to find any inherent grouping that are already present in the data, can be done via unsupervised learning. By developing rules based on the data and identifying connections or links between them, it can also be utilized to solve association problems. The semi-supervised machine learning

uses both labelled and unlabeled data for training, it sits in between supervised and unsupervised learning. Typically, semi-supervised machine learning uses a small amount of labelled data and a significant amount of unlabeled data. This technique enables systems to significantly increase learning accuracy.

C. Machine Learning Tools

Machine learning models can be created using a wide variety of software tools, and they can then be applied to fresh, unexplored data. Additionally, a sizable selection of well described machine learning algorithms are readily available. Most of the time, these tools include libraries that implement some of the most well- liked machine learning techniques. They fall into the following categories:

Solutions based on pre-built applications. Programming languages with dedicated machine learning libraries We had more control over the algorithmic parameters when we developed and implemented models using programming languages since it is more flexible. It also enables us to comprehend the output models created better. In the area of machine learning, some common programming languages include:

Python: Python is a very well-liked option in the development of machine learning and AI. It is quite simple to learn thanks to its short and straightforward syntax.

R: R is one of the best and most productive languages for statistical data manipulation and analysis. We can quickly create publication-quality plots using R by adding mathematical notation and formulas as necessary. R is a general-purpose language, but it also offers a tonne of machine learning-related packages including RODBC, Gmodels, Class, and Tm.These programmes facilitate the use of machine learning techniques for solving business-related problems.

Tensor Flow: It is an open source, end-to-end machine learning platform. Researchers can advance the state-of-the-art in machine learning thanks to its extensive, adaptable ecosystem of tools, libraries, and community resources, while developers can simply create and deploy ML-powered applications.TensorFlow offers reliable Python and C++ APIs as well as backward compatibility that is not guaranteed.

D. Machine Learning algorithms used in the recommendation system are:

Logistic Regression: When the dependent variable is dichotomous (binary), logistic regression is the suitable regression approach to do. The logistic model (also known as the logit model) is used to simulate the likelihood that a class or event, such as pass/fail, win/lose, alive/dead, or healthy/ill, will occur. This can be expanded to simulate a variety of event classes, such identifying the presence of a cat, dog, lion, etc. in an image. Each object in the image that is detected would be given a probability between 0 and 1, with the sum adding up to 1.

Neural Networks: These are a class of algorithms that are made to recognize patterns and are loosely based on the human brain. They categorize or group raw input to understand sensory data using a form of machine perception. All real- world data, including images, sounds, texts, and time series, must be converted into vectors for them to recognize the patterns, which are numerical and contained therein. We can classify and cluster data using neural networks. Since neural networks inherently approximate general functions, they can be used to solve virtually any machine learning problem involving learning a complex mapping from the input to the output space.

Decision Tree: Regression and classification issues using a decision tree. By generating decision rules from training data sets, decision trees can be used to build models that predict classes or values of target variables. The decision tree algorithm mimics the root, branch, and leaf structure of a tree. Class labels are represented as

leaf nodes, whereas decision-making attributes are internal nodes. Compared to other classification methods, the Decision Tree approach is simple to understand.

NN (K-nearest Neighbour): The k-nearest neighbours algorithm (kNN) is a non-parametric technique used for classification and regression in pattern recognition. The k closest training instances in the feature space make up the input in both scenarios. Whether k-NN is applied for regression or classification determines the results. The result of k-NN classification is a class membership. An object is allocated to the class that has the most support from its k closest neighbours (k is a positive integer that is often small) based on a majority vote of those neighbours. The object is simply put into the class of its one nearest neighbour if k = 1. The output of a K-NN regression is the object's property value. The average of the values of the k closest neighbours makes up this number. K-NN is a form of instance-based learning, often known as lazy learning, in which all computation is postponed until after the function has been evaluated.

Gaussian NB: Nave Bayes classifiers are a family of straightforward "probabilistic classifiers" used in machine learning. They are based on applying the Bayes theorem with strong (nave)independence assumptions between the features. One of the simplest Bayesian network models is this one. They might, however, be combined with kernel density estimation to attain better levels of accuracy. A common presumption when working with continuous data is that the continuous values connected to each class are distributed in accordance with a normal (or Gaussian) distribution.

IV. RESULTS AND DISCUSSION

Analysis of Data: Analyzing the data is one of the initial actions we take throughout deployment. We did this in an effort to determine whether there were any connections between the different attributes that were available in the dataset. Getting a training data set: The UCI Repository for renal disease provided the Training data set. The values in the dataset are actual test results that were received, and the dataset was gathered from several hospitals in Tamil Nadu. The dataset has a total of 24 properties, but preprocessing revealed that just 6 of them are crucial for establishing that association. There are 400 samples in all.

Preprocessing: Data preprocessing comes after data analysis and visualization. Preprocessing the data is a crucial step since it helps to clean the data and prepare it for usage in machine learning algorithms. Preprocessing primarily focuses on addressing any missing values and outliers as well as inaccurate or outlier-containing data. There are two approaches to deal with missing data. The first approach is to just eliminate the entire row that contains the inaccurate or missing value. Although this method is simple to apply, it is best to limit its application to huge datasets. This strategy can result in an excessive reduction in the size of tiny datasets, especially if there are many missing values. The accuracy of the outcome may suffer significantly as a result. We won't be employing this strategy because the dataset we have is relatively tiny. Instead, depending on the type of attribute, we would substitute the average or mode of the column for the missing values. If the attribute is non-nominal, we would use the mode; if it is nominal, we would use the average. Since the values in the dataset we utilized were in string format, we had to transform and encode them into integer values before feeding them to the neural network. The data was first transformed into pandas categorical data, and distinct data frames were created.

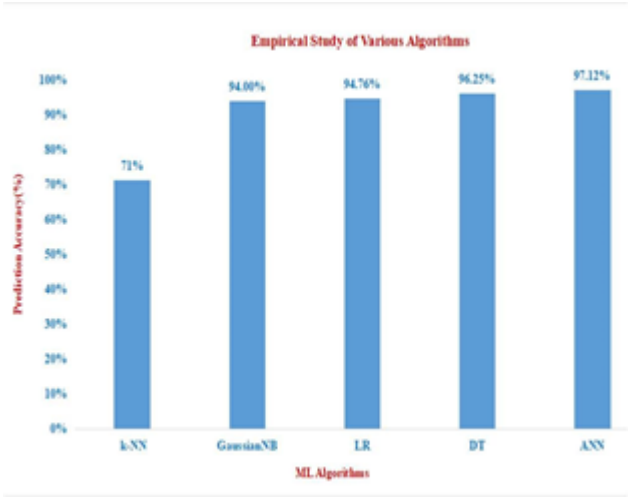

Fig 1. Comparative Analysis of Various ML Algorithms

Figure 1 depicts the comparative analysis of numerous machine learning algorithms. The research study used a range of evaluation approaches to examine the models, which improves the accuracy of case diagnosis. Because the proposed approach is simple to use, the adoption of such a framework is also feasible. Different prediction models, including K-Nearest Neighbour Classifier, Logistic Regression, Decision Tree Classifier, and Artificial Neural Network (ANN), were trained on the processed dataset. The models' performance was predicted to exhibit higher accuracy and significance.

V. CONCLUSION

The best prediction method to identify CKD at an early stage was presented by this system. The models are trained and validated for the provided input parameters, and the dataset displays input parameters gathered from CKD patients. The diagnosis of CKD is carried out using the K-Nearest Neighbours Classifier, Decision Tree Classifier, GaussianNB, Logical Regression, and Artificial Neural Network learning models. Accuracy, Specificity, Sensitivity, and Log Loss are some of the comparative measures that are used to assess the performance of the models. The study's findings demonstrated that, when all metrics are taken into account, the Logical Regression model outperforms all other models in its ability to predict CKD. This technique would make it possible to predict a person's likelihood of developing CKD later in life, which would be extremely beneficial and affordable for most people. If a person is at risk, this model might be coupled with standard blood report creation to instantly flag that person. Patients wouldn't need to visit a doctor until the algorithms flagged them. For the current busy person, it would be less expensive and simpler as a result.

VI. ACKNOWLEDGMENT

I am thankful for the opportunity to work on this and other relevant projects with everyone. Each of the members of my Dissertation Committee has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general.

The people in my family have been more significant to me in the pursuit of this purpose than anyone else. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

VII. REFERENCES

[1]. Manish Kumar, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, pg. 24-33, 2016.

[2]. Nayak, Lipsa & Dharmarajan, K, "A Survey on Chronic Kidney Disease Detection Using Novel Methods" , 2018.

[3]. Xiao, J.; Ding, R.; Xu, X.; Guan, H.; Feng, X.; Sun, T.; Zhu, S.; Ye, Z. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J. Transl. Med. 2019, pp. 17, 119.

[4]. Ghosh, P.; Shamrat, F.J.M.; Shultana, S.; Afrin, S.; Anjum, A.A.; Khan, A.A. Optimization of prediction method of chronic kidney disease using machine learning algorithm. In Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 18– 20 November 2020; pp. 1–6.

[5]. Ifraz, G.M.; Rashid, M.H.; Tazin, T.; Bourouis, S.; Khan, M.M. Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods. Comput. Math. Methods Med. 2021, 2021, 6141470.

[6]. CKD Prediction Dataset. Available online: https://www.kaggle.com/datasets/abhia19 99/chronic-kidney-disease (accessed on 27 June 2022).

[7]. Islam, M.A.; Akter, S.; Hossen, M.S.; Keya, S.A.; Tisha, S.A.; Hossain, S. Risk factor prediction of chronic kidney disease based on machine learning algorithms. In Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 3– 5 December 2020; pp. 952–957.

[8]. Yashfi, S.Y.; Islam, M.A.; Sakib, N.; Islam, T.; Shahbaaz, M.; Pantho, S.S. Risk prediction of chronic kidney disease using machine learning algorithms. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020; pp. 1–5.

[9]. Chittora, P.; Chaurasia, S.; Chakrabarti, P.; Kumawat, G.; Chakrabarti, T.; Leonowicz, Z.; Jasiński, M.; Jasiński, Ł.; Gono, R.; Jasińska, E.; et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access 2021, 9, 17312–17334.

[10]. Revathy, S.; Bharathi, B.; Jeyanthi, P.; Ramesh, M. Chronic kidney disease prediction using machine learning models. Int. J. Eng. Adv. Technol. (IJEAT) 2019, 9, 6364–6367.

[11]. Yadav, D.C.; Pal, S. Performance based Evaluation of Algorithmson Chronic Kidney Disease using Hybrid Ensemble Model in Machine Learning. Biomed. Pharmacol. J. 2021, 14, 1633– 1646.

[12]. Baidya, D.; Umaima, U.; Islam, M.N.; Shamrat, F.J.M.; Pramanik, A.; Rahman, M.S. A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1305–1310.