Comparison Between Various Machine Learning Classification Techinques in Detecting Chronic Kidney Disease

Rakib Hossen
Department of Biomedical Engineering
Bangladesh University of Engineering and Technology
Dhaka,Bangladesh
E-mail: rakibbuet@outlook.com

Shakib Mahmud Ayon
Department of Biomedical Engineering
Bangladesh University of Engineering and Technology
Jashore, Bangladesh
E-mail: shakib.mahmud.ayon@gmail.com

***Abstract*—Application of Artificial Intelligence and related fields are increasing day by, rather being confined to the core areas of computer science, they are spreading into various new domains. In recent times, Machine Learning i.e. a sub-domain of AI has been widely used in order to assist medical experts and doctors in the prediction, diagnosis and prognosis of various diseases and other medical disorders. In this manuscript, the authors applied various machine learning algorithms to a problem in the medical diagnosis domain and analyzed their performance efficiency in predicting the results. The selected problem is the diagnosis of Chronic Kidney Disease. The used dataset consists of 400 instances and 24 attributes. The authors evaluated 10 classification algorithms by applying them to the Chronic Kidney Disease Dataset. The prediction by the candidates were compared with the actual medical results of the subject to calculate the candidates' efficiency. The used performance evaluation metrics are predictive accuracy, specificity, sensitivity and precision. The results indicate that RF performed best with 100%, specificity of 1, sensitivity of 1 and precision of 1.**

***Index Terms*—Artificial Intelligence, AI, Machine Learning,Chronic kidney disease, Medical diagnosis, Decision-Tree, Support vector machine, SVM, Artificial neural networks, ANN, K-nearest neighbour, KNN.**

I. Introduction

AI refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions [1]. Computer systems, that simulate human intelligence processes: learning, reasoning and self-correction are called artificial intelligence systems [2]. Machine learning, a branch of AI, is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Nowadays machine learning is used as a connector between computer science, applied mathematics and statistics sharing concepts with AI and information theory. AI programs are hugely applied to practices in the medical field such as diagnosis processes, treatment protocol development, drug development, personalized medicine, and patient monitoring and care. AI algorithms can also be used to analyze large amounts of data through electronic health records for finding out patterns from the training data which can be used later for disease prevention and diagnosis.

Chronic kidney disease (CKD) is a condition characterized by a gradual loss of kidney function over time. The disease is called "chronic" because the damage to the kidneys happens slowly over a long period of time. There is a high mortality rate due to chronic kidney disease. If kidney disease gets worse, wastes can build to high levels in blood. There might be complications like high blood pressure, anemia (low blood count), weak bones, infection, cognitive impairment, poor nutritional health and nerve damage. Also, it increases the risk of having heart and blood vessel disease [3]. It can get worse over time and eventually the kidneys may stop working altogether which might lead to death if not treated properly. When the dysfunctionalities of the kidneys increase at an alarming rate, dialysis and transplantation are used for treatment [4]. Among all the five stages of chronic kidney disease this is known as the end-stage [5].

We have organized the rest of the paper in the following order: The IInd section represents the data used and various candidate techniques. Related works have been represented in IIIrd section. In IVth section, implementation of the methodology and results have been analyzed and discussed. Finally, in the Vth section we have concluded the paper and described future works.

II. Materials and Methods

A. Chronic kidney disease data

The dataset utilized to conduct the study in this paper was obtained from the University of California, Irvine data repository. A total of 400 persons from the southern portion of the country are included in the dataset. of India, with ages ranging from 2 to 90 years. There are a total of twenty-four criteria, the majority of which are clinical in nature. The rest is physiological, as is nature. Table 1 summarizes the findings. Missing values were handled with during data preparation by substituting numeric and integer values by attribute mean of the all the instances with consideration and nominal values were replaced using attribute mode.

B. Candidate classification techniques

Ten different classifiers majorly based on the following techniques: Decision-Tree, Support vector machine (SVM), Random Forest, K-nearest neighbor (KNN), Artificial neural networks (ANN) were analyzed. Because of popularity in recent literature, these algorithms were selected. A brief description about the selected techniques has been given below:

- Support Vector machine(SVM): Support Vector Machine is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. SVM finds a hyper-plane that creates a boundary between the types of data. In two-dimensional space, this hyper-plane is nothing but a line. The underlying algorithm is Support Vector Classification (SVC) and it revolves around the perception of a "margin"- on either side of a hyperplane that divides two data classes. Maximizing the margin creates the largest possible distance among the hyperplane and the instances on either side of the hyperplane reduce an upper bound on the anticipated generalization error. It works on two types of data i.e. linearly separable data and nonlinearly separable data. SVM works very well without any modifications for linearly separable data. Linearly separable data is any data that can be plotted in a graph and can be separated into classes using a straight line. And for nonlinearly separable data kernelized SVM can be used. A kernel is nothing but a measure of similarity between data point.

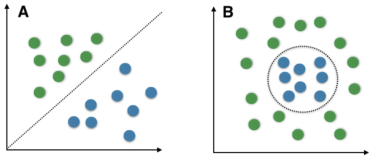

Fig. 1. A: Linearly Separable Data B: Non-Linearly Separable Data

- Decision tree: The decision tree model can forecast both categorical and continuous variables. It, like SVM, may be used for regression or ranking. As a result, trees are classified into two types: classification decision trees and regression decision trees. The construction of a decision tree is straightforward: it hierarchically divides data into subsets, which are then divided further into smaller partitions or branches until they become "pure", which means that the characteristics within the branch all belong to the same class. These classifications are referred to as "leaves". There are number of ways for finding the feature that best divides the training data, some of them are namely Information gain, myopic measures, G-statistics, chi-square, MDL, etc. [6]

- K-Nearest Neighbors (KNN): This is the simplest of all machine learning algorithm. The most basic form of the K-nearest neighbor classifier algorithms predicts the target label by locating the nearest neighbor class. The closest class will be identified using the distance measures like Euclidean distance [7]. It is a non-parametric algorithm which means that it does not assume anything on the underlying data distribution. The Euclidean distance between the test data and each sample in the training data is computed, followed by categorizing the test data into a class to which the majority of the k-closest neighbors of the training data belong. Determining the value of K in K-nearest neighbor is the most difficult thing to do. A lower number of K indicates that noise will have a greater impact on the result, implying that the chance of overfitting is quite high. A large value of K makes it computationally costly, which contradicts the primary purpose of KNN.

- Artificial Neural Network(ANN): The Artificial Neural Network is a deep learning algorithm that originated from the concept of Biological Neural Networks in human brains. The development of ANN resulted from an attempt to replicate the workings of the human brain. ANN functions in a way that is similar to but not identical to biological neural networks. It is a sophisticated data modeling tool that can capture, describe, and simulate complicated connections between inputs and outputs using numerous concurrent calculations. Traditional Machine Learning algorithms tend to perform at the same level when the data size increases but ANN outperforms traditional Machine Learning algorithms when the data size is huge. A neural network with a single layer is called a perceptron. A multi-layer perceptron is called Artificial Neural Networks.
  A neural network can possess any number of layers. Each layer can have one or more neurons or units. Each of the neurons is interconnected with each and every other neuron. Each layer could have different activation functions as well. Figure 2 illustrates a typical artificial neuron.

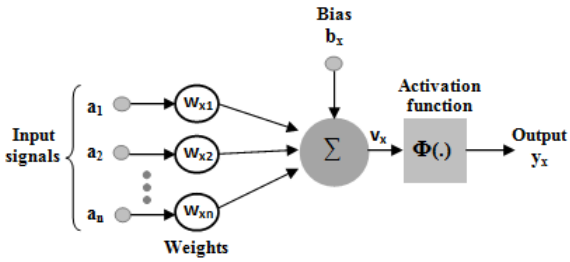

Fig. 2. An artificial neuron

C. Methodology

After complete preprocessing, all attributes corresponding to each case are described initialy i.e. input data and out of

these only one attribute is used to represent a decision for the given problem i.e. output data. For every individual classifier 5-fold cross validation is used. The procedure is:

- Partitioning the data into k folds.
- For all folds from cross validation score mean of accuracy, sensitivity, precision and specificity are calculated.
- All scores were compared in a bar plot.
- Excluding three columns with largest number of missing values first twosteps were done, but only for KNN, Fine KNN, Linear SVM, Decision Tree and Random forest classifiers. Then the scores were compared with similar classifier scores of whole dataset.

III. Related Works

Machine Learning and data-mining is used by a number of researchers to solve different problems in the field of medicine. Sahil Sharma et. al. [8] applied 12 different machine learning algorithms to diagnose Chronic Kidney Disease and analyzed their efficiency in predicting the results. The results indicate that decision-tree performed best with nearly the accuracy of 98.6%, sensitivity of 0.9720, precision of 1 and specificity of 1.

Abid Sarwar et al. [9] compared the accuracy of different machine learning algorithms, such as- Naïve Bayes, artificial neural network, and KNN, for the type II diabetes. The condition when the pancreas is not able to produce the needed amount of insulin or the cell is not able to use the produced insulin (insulin resistance) which leads to abnormal glucose level in the blood is termed as a type II diabetes. Prediction accuracy in this comparison study shows that Neural network with 96% prediction accuracy performs better than Naïve Bayes with 95% and KNN 91%.

Igor Kononenko [10] presented a view on the use of Machine learning techniques-

1. For the interpretation of medical data,
2. For intelligent analysis of medical data in the current scenario and
3. For assistance of physicians in diagnosis of medical disorders, in the future.

Integration of machine learning techniques with the existing instrumentations is suggested by the authors for the acceptance of machine learning in medicine.

Yasodha et al. [11] analyzed a database of diabetic patients using weka tool considering different algorithms such as REP Tree, Bayes Network, J48 and Random Tree classifiers for the study and compared the outputs. The main objective of the study was to develop a Diabetic expert system; inputs being patients' daily glucose rate and insulin dosages, the system would predict the patients' insulin dosage for the next day.

T.Manju et al. [12] proposed a hybrid system of multi-layer feed forward neural network (MLFFN) and genetic algorithm for assisting doctors in predicting the heart disease. The heart attack is the major cause of death in the world; its major causes are smoking, high blood pressure, unhealthy diet, obesity and diabetes.

The ANN is trained using back propagation and feed forward neural network. Genetic algorithm is used for the weight optimization. The weights are associated with each connection in the neural network nodes. The data set consisted of 13 factors out of which only 6 were used for the purpose of training the ANN. The ANN then predicted the possibility of heart attack in a patient or not.

Babak Sokouti et.al. [13] proposed Levenberg–Marquardt (LM) feedforward MLP neural network in order to classify cervical cell images obtained from 100 patients including healthy, low-grade intraepithelial squamous lesion and high grade intraepithelial squamous lesion cases. The semi-automated cervical cancer diagnosis system is composed of two phases: image pre-processing/processing and feed forward MLP neural network. The results showed that cervical cell images were classified successfully with 100% correct classification rate using the proposed method.

IV. Implementation and Results

All the classifiers were applied to the same dataset using Google Colab and the results were obtained and analyzed in the terms of predictive accuracy, sensitivity, precision & specificity. Z% predictive accuracy shows that the classifier is able to classify nearly Z% of instances correctly. Sensitivity can be defined as the ability to correctly detect cases which do have the condition. Mathematically it can be defined as the number of true positives divided by the number of true positives and false negatives. Specificity can be defined as the ability to correctly detect cases which do not have the condition. Mathematically, it can be expressed as the number of true negatives divided by the number of true negatives and false positives. Precision is the number of correctly classified positive examples divided by the number of examples labeled by the system as positive [14] i.e.total number of true positives divided by the total number of true positives divided by the total number of true positives and false positives. True positives are the correctly recognized class examples, true negatives are correctly recognized examples that do not belong to the class, whereas false positives are examples that were incorrectly assigned to the class and false negatives are those examples that were not recognized as class examples [14]. The maximum value of Sensitivity, Precision, and Specificity can be 1 whereas their minimum value can be 0. The higher the value, the better the performance of the classifier considered.

From Fig. 3, we can deduce that RF algorithm shows the best classification result with scoring 100% in accuracy, 1 in specificity, 1 in sensitivity and 1 in precision. This

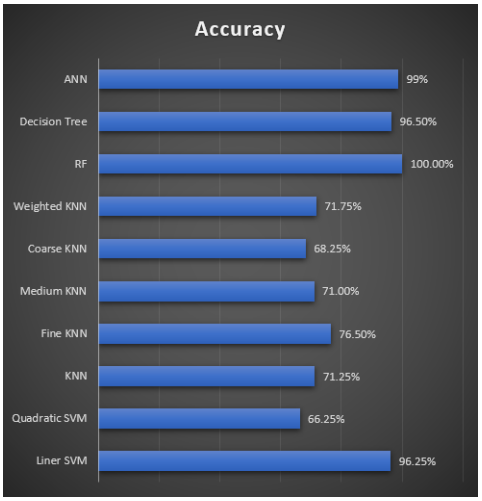

Fig. 3. Comparison of predictive accuracy

algorithm performed better than all other classifiers in all four-performance metrics.

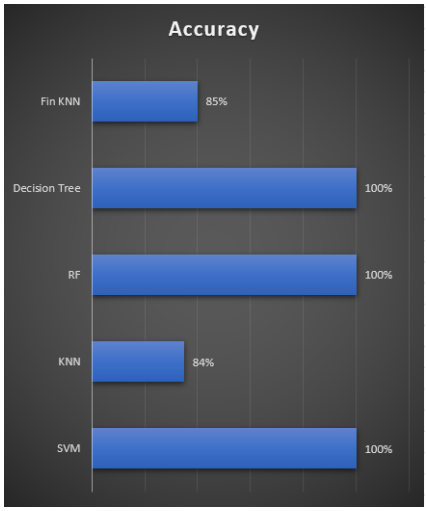

Fig. 4. Comparison of predictive accuracy excluding attributes with over 100 missing data

For Fig. 4, we excluded attributes such as RBC, RBC count and WBC count. We excluded those attributes as they had more than 100 missing values in the provided dataset. This time the classification accuracy changed drastically. We can see that RF performs similar accuracy score while Decision Tree and SVM improves to 100%. Default KNN and Fine KNN also improves from the previous accuracy score.

We can see the values of specificity, sensitivity and precision of the candidates in Table I. And Fig. 5 lllustrates comparative values of Table I.

In Table II, we can see the values of Fine KNN, Decision Tree, RF, KNN and SVM and in Fig. 6, the table is

TABLE I
SPECIFICITY, SENSITIVITY AND PRECISION VALUES OF ALL CANDIDATES

Classifier Name	Specificity	Sensitivity	Precision
Linear SVM	0.964	0.962	0.9585
Quadratic SVM	0.7719	0.626	0.636
KNN	0.6559	0.7313	0.7196
Fine KNN	0.752	0.7693	0.7558
Medium KNN	0.6839	0.7186	0.7089
Coarse KNN	0.72	0.67	0.6671
Weighted KNN	0.648	0.7406	0.7295
Decision Tree	0.972	0.9626	0.9631
RF	1	1	1
ANN	0.9787	1	0.9706

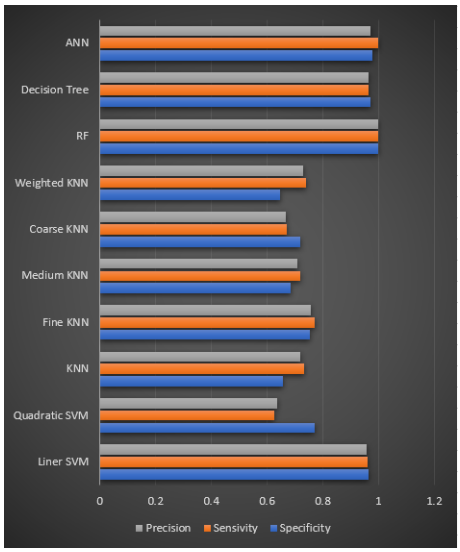

Fig. 5. Comparison of specificity, sensitivity and precision

representated in a chart.

The excluded parameters bear not too much importance according to the chart in Fig. 7.

V. CONCLUSION AND FUTURE SCOPE

Chronic kidney disease means a gradual loss of kidney function over time. 1 in 10 people worldwide have chronic kidney disease, according to The Global Kidney Health Atlas. 1 in 3 people in the general population worldwide is at increased risk of chronic kidney disease, and the report

TABLE II
SPECIFICITY, SENSITIVITY AND PRECISION VALUES OF ALL CANDIDATES EXCLUDING ATTRIBUTES WITH OVER 100 MISSING DATA

Classifier Name	Specificity	Sensitivity	Precision
Linear SVM	1	1	1
KNN	0.832	0.8393	0.8295
RF	1	1	1
Decision Tree	1	1	1
Fine KNN	0.868	0.844	0.8398

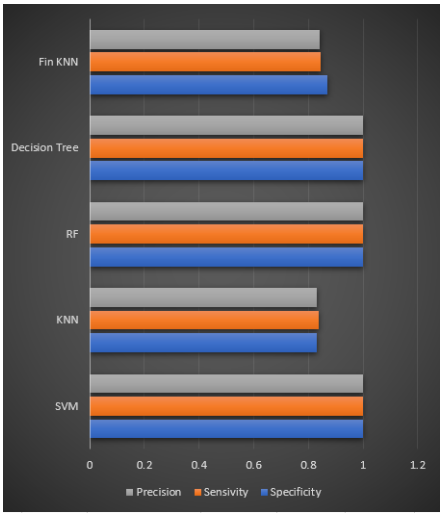

Fig. 6. Comparison of predictive accuracy excluding attributes with over 100 missing data

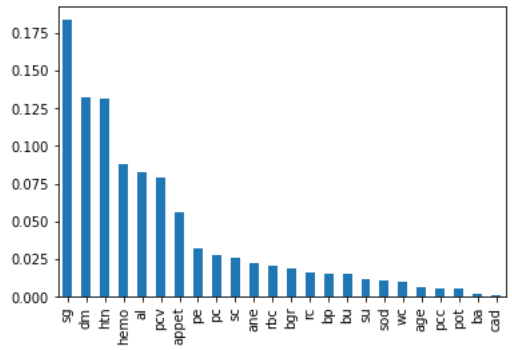

Fig. 7. Feature importance graph

estimates that 9 in 10 of those with kidney disease are unaware of their condition [15]. In underdeveloped and developing countries, people have a higher percentage of chronic kidney disease cases than in developed countries. This is because most of the cases in underdeveloped and developing countries remain undiagnosed or are diagnosed later due to the lack of routine check-ups in these countries. A report says: "Patients who receive treatment for kidney failure worldwide, more than 80% of them are in developed countries with large elderly population and access to health care." [16]. Machine learning-based programs can be used for the cases of underdeveloped and developing countries for timely and accurate detection by using prediction from collected data. This will be helpful for the doctors as they will not have to go through the whole diagnosis process manually. They will be able to verify the results of diagnosis in less time and will be able to attend more patients at the same time.

For future works, one can implement some hybrid or ensemble techniques to improve the two-class classification accuracy. Also, one can see the effect on performance by extracting different subsets of features from the complete dataset. If accuracy remains unchanged or close to the previous after extraction, then the extracted features can be taken out from the list of clinical tests and this will reduce the financial burden of the patients as they will have to undergo a smaller number of tests than before.

REFERENCES

[1] Jake Frankenfield, "Artificial Intelligence (AI)", Investopedia, 2021.

[2] "What is Artificial Intelligence System (AIS)", IGI Global.

[3] "Chronic Kidney Disease (CKD) Symptoms and causes", NATIONAL KIDNEY FOUNDATION, 2017.

[4] "Chronic kidney disease (CKD)", American Kidney Fund, 2021.

[5] "The Stages of Kidney Disease", FRESENIUS KIDNEY CARE.

[6] Olga Berezovsky, "The decision tree classifier – An overview", LOGIC2020, 2020.

[7] Rahul Saxena, "KNN CLASSIFIER, INTRODUCTION TO K-NEAREST NEIGHBOR ALGORITHM", Dataaspirant, 2016.

[8] A. S. a. V. S. Sahil Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," Springer Link, 2016.

[9] V. S. Abid Sarwar, "Comparative analysis of machine learning techniques in prognosis of type II diabetes," AI & Society, 2013.

[10] I. Kononenko, "Machine Learning for Medical Diagnosis:," Artificial Intelligence in, vol. Vol.23, no. 1, 2011.

[11] M. K. P Yasodha, "Analysis of Population of Diabetic," International Journal of Science, vol. 2, no. 5, 2011.

[12] K. P. a. R. T.Manju, "Heart Disease Prediction System using Weight Optimized Neural Network," International Journal of Engineering and Computer Science, Vols. Vol- 2, no. No.3, 2013.

[13] S. H. a. A. D. T. B. Sokouti, "A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features," Neural Comput. Appl., vol. Vol. 24, no. No. 1, p. pp. 221–232, 2014.

[14] G. L. Marina Sokolova, "A systematic analysis of Performance measures for classification tasks," Information Processing and Management, Elsevier, Vols. Vol.-45, no. Issue-4, pp. pp.427-437, 2009.

[15] "1 in 10 people worldwide have CKD, but many governments don't make kidney disease a priority", Healio, 2017.

[16] V Jha, G Garcia-Garcia, K. Iseki, et.al. "Chronic kidney disease: global dimension and perspectives "Lancet. , pp. 260-272 , 2013.

ACKNOWLEDGMENT

Sumaiya Ohab (S201718003), Md. Rakib Hossen (S201718011), Shakib Mahmud Ayon (S201718023) are currently pursuing their BSc in Biomedical Engineering (BME) from Department of Biomedical Engineering, Bangladesh University of Engineering and Technology(BUET), Dhaka, Bangladesh.