IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

HYBRID MACHINE LEARNING MODEL FOR DIAGNOSIS OF CHRONIC KIDNEY DISEASE WITH OPTIMAL FEATURE SELECTION

[1]C.KARTHIK, [2]A. THULASI, [3]G. MRUDULA, [4]A. SREEJA,
[5]D. CHANDANA LIKITHA, [6]A.L.S. HARSHITHA, [7]A. THEJASWINI
[1]Associate Professor, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru(M), Nellore (DT), Andhra Pradesh, India.
[2,3,4,5,6,7]B.Tech Scholars, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru (M), Nellore (DT), Andhra Pradesh, India.

ABSTRACT: Currently, there are many people are suffering from chronic kidney diseases worldwide. Classification of kidney disease is vital for global improvement and accomplishment of practical guidance. Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. This paper presents, Hybrid Machine Learning model for diagnosis of Chronic Kidney Disease (CKD) with optimal feature selection. The patient with CKD and non-CKD status can be predicted using hybrid machine learning classification algorithms. Used machine learning classifications in this hybrid model are Naïve Bayes (NB), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The experiments are applied to the UCI Machine Learning Repository dataset. The proposed method selects applicable features of kidney data with the help of Ant Lion Optimization (ALO) technique to choose optimal features for the classification process. Performance comparison indicates that our proposed model accomplishes better classification accuracy, precision, F-measure, sensitivity measures when compared with other classifiers.

KEYWORDS: Chronic Kidney Disease (CKD), Ant Lion Optimization, KNN, SVM, NB, Machine Learning.

I. INTRODUCTION

Chronic diseases have become the newest threat to the developing nations. Based on World Health Organization (WHO), chronic disease cases increase rapidly in developing countries and are becoming a major concern all over the world [1]. Most of the preventive measures have mainly focused on (CVD). However, the CKD is considered one of the chronic diseases due to the increased number cases and its consequences.

Chronic Kidney Disease (CKD) is one of the types of kidney disease, which results in a gradual loss of kidney function. This phenomenon can be observed over a period of months or years due to several living conditions of patients [2]. The CKD is also called a chronic kidney failure where according current medical statistics the 10% of the population worldwide is affected by CKD. According to the World Health Organization (WHO) 35 million attributed to chronic diseases. Currently it is estimated that one in five men, and one in four women aged 65 through 74 are going to be affected by CKD worldwide. CKD is unique in its nature among most diseases since it is mostly discovered when it is in the final stages of progression whereby it will be much risky as well as expensive to treat due to being in the final stage called kidney failure. The final stage of chronic kidney disease is called end-stage renal disease (ESRD). At this stage, the kidneys are no longer able to remove enough wastes and excess fluids from the body. The patient needs dialysis or a kidney transplant [3].

Kidneys are two bean-shaped organs, each about the size of a fist. They are located just below the rib cage, one on each side of the spine. Every day, the kidneys filter about 120 to 150 quarts of blood to produce about 1 to 2 quarts of urine. The key function of the kidneys is to remove waste products and excess fluid from the body through the urine. The production of urine involves highly complex steps of excretion and re-absorption. This process is necessary to maintain a stable balance of body chemicals. The critical regulation of the body's salt, potassium and acid content is performed by the kidneys and produce hormones that affect the function of other organs. For example, a hormone produced by the kidneys stimulates red blood cell production, regulate blood pressure and control calcium metabolism etc.

Diagnosing CDK usually starts with clinical data, lab tests, imaging studies and finally biopsy. Although biopsy is the standard diagnosing test, it has many disadvantages, such as being invasive, costly, time-consuming and sometimes risky. For example; when a biopsy is performed, the patient may face infection, the scare of surgery and misdiagnosis [4]. Imaging studies (mammogram, sonogram, and MRI of the kidney) has been used for many years to detect the disease. But using them has some limitation; more expressly is exposure effects of radiation. Besides being risky, the data provided by imaging is insufficient to diagnose CDK. The automated diagnosis of different diseases has attracted many researchers.

Diagnosis of most diseases has heavy cost since many experiments required to predict the disease. Selecting attributes which are really important for prediction of disease can reduced this cost. Thus dimensionality reduction plays an important role in medical diagnosis. Some recent studies which widely use feature selection techniques are diagnosis of breast cancer, erythemato-squamous diseases, and CT focal liver lesions. Machine learning is being used to intelligently interpret available data and transform it into useful knowledge to increase the diagnostic process efficiency. Machine learning is already being used to assess the state of the human body, analyze disease-related aspects, and diagnose a variety of disorders. CKD therapy concentrates on minimizing the movement of kidney risk by regulating the basic reason for the disease at initial stages.

II. LITERATURE SURVEY

Muhsen, Heba, et al. [5] proposed a system that combines deep learning and discrete wavelet transform features techniques. They used fuzzy c-mean method to segment the brain tumor, and for every detected lesion the wavelet transform features was used to extract the features, then these features are fed into the principal component analysis for feature dimension reduction and lastly the selected features are then fed to deep neural networks. The results showed that they achieved an accuracy rate of 96.96% and a sensitivity rate of 97.1 %.

Widhiarso, Wijang, Yohannes Yohannes, and Cendy Prakarsah. et al. [6] proposed a brain tumor classification system by using convolutional neural network and Gray Level Co-occurrence Matrix (GLCM) based features. They extracted four features (Correlation, Contrast, Energy, and Homogeneity) using four different angles (0°, 45°, 90°, and 135°) for every image and these features are supplied into the CNN, they evaluated the proposed system using four different datasets (GlPt, Mg-Gl, Mg-Pt, and Mg-Gl-Pt) and the highest accuracy obtained was 82.27% for Gl-Pt dataset using two sets of features; contrast with homogeneity and contrast with correlation. Khawaldeh, Saed, et al. [7] proposed a system for non-invasive classification of glioma brain tumors using a modified version of AlexNet CNN. The classification process was performed by dint of MRI images of the entire brain and the labels of the images were at the image level, not the pixel level. The result of evaluating the system showed an accuracy of 91.15%.

Vijayarani, S., and S. Dhayanand. et. al. [8] the main objective of this research is to model kidney disease using structure algorithms such as SVM and NB. This exploration work for the most part centered around finding the best grouping algorithm dependent on the characterization precision and execution time execution factors. K.R Lakshmi, Y. Nagesh, and M. VeeraKrishna, et. al. [9] uses three supervised machine learning algorithms i.e., Decision trees (DT), Logical Regression (LR) and Artificial Neural Networks (ANN) performed classification for Kidney dialysis data. The tool named Tanagra used to perform the classification. For the classifiers evaluation, the 10-fold cross validation is used. The experimental results showed that ANN outperformed by 93.8% remaining algorithms.

G. Caocci, R. Baccoli, R. Littera, S. Orrù, C. Carcassi and G. La Nasa, et. al. [10], tried to predict the Long Term Kidney Transplantation Outcome. They have performed comparative analysis between an ANN and LR algorithms. The comparative analysis has been implemented based on performance metrics like accuracy, sensitivity and specificity. During the study for the kidney transplant recipients prediction of kidney rejection which was based on ten training and validating datasets. The experimental results showed that, ANN can be considered a useful supportive algorithm in the prediction process of the defined problem. In summary, the ability of predicting kidney rejection (sensitivity) was 38% for LR versus 62% for ANN. The ability of predicting no-rejection (specificity) was 68% for LR compared to 85% of ANN.

III. HYBRID MACHINE LEARNING MODEL FOR DIAGNOSIS OF CHRONIC KIDNEY DISEASE

The block diagram of Hybrid Machine Learning model for diagnosis of Chronic Kidney Disease (CKD) with optimal feature selection is represented in below Fig. 1.

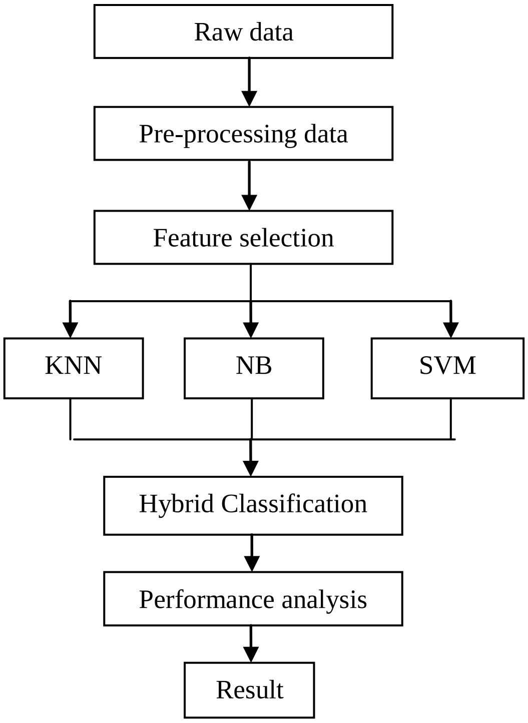

Fig. 1: BLOCK DIAGRAM OF DIAGNOSIS OF CHRONIC KIDNEY DISEASE

The experiments are applied to the UCI Machine Learning Repository dataset. The detection of CKD is based on 23 attributes. The dataset has 400 data instances of which 250 are positive CKD and 150 are negative CKD.

The dataset we received from the internet source needs to be cleaned as it has null or NA values for various attributes for a given instance. Missing values in the dataset like NA's or blank values are removed by using Pandas library "dropna" and "fillna", which drops either column or rows with missing data and replaces NA's with the mean values of that attribute respectively.

Feature selection and dimension reduction are fundamental walks in pattern recognition errands. In this examination, regardless of the way that the feature set was not outrageous and attaining appealing outcomes, using the majority helpful features extended the classification rate. For removing a few features, the existing work is observed to be scattered which assesses the factorization, probability and entity connections. Our proposed model picks the ideal features using Ant Lion Optimization (ALO). Initially, the optimal features are selected with the help of this algorithm and provide the optimal solution.

Antlions (doodlebugs) have a place with a class of netwinged insects. ALO is inspired by the food searching behavior of antlions. In the wake of burrowing the trap, larvae placed below the base of the cone-shaped trap & sit firmly for insects (ideally ant) to be wedged in the pit. The edging of the trap is amply pointed for insects to fall down to the bottom of the trap effortlessly.

After feature selection, the data is pass through machine learning classifiers. The combination of three classifiers is used in this paper which is called as hybrid model. These three classifiers are SVM (Support Vector Machine), and Naïve Bayes (NB). The SVM is a supervised learning algorithm that is used for data classification and regression. It searches for a best hyperplane which separate between classes. The best hyperplane is considered the one which leaves the maximum margin between the two distinct classes. The margin is defined as the width of the hyperplane from the closest point of the two distinct classes. Bounds between data sets and hyperplane are called support vectors.

Naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with independence assumptions. In other words, this probability model would be an 'independent feature model'. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other features. Naive Bayes classifier performs reasonably well even if the underlying assumption is not true.

K-nearest neighbors (KNN) algorithm belongs to the instance-based methods, which simply store all the training examples and delay the classification task until a new instance must be classified. The classification task is as follows: first it finds the set of k nearest neighbors to n, where n is a new example to classify. Then the algorithm takes the plurality vote of the neighbors.

After the classification of all these four type of classifiers, all results are combined in ensemble module. Then the exact output or results are evaluated by using performance metrics such as Accuracy and Precision.

IV. RESULT ANALYSIS

The experiments are applied to the UCI Machine Learning Repository dataset. The detection of CKD is based on 23 attributes.

The dataset has 400 data instances of which 250 are positive CKD and 150 are negative CKD. 75% of total dataset is used as training and remaining 25% of data is used for testing. Accuracy, Sensitivity, Precision, and specificity are used parameters and computed as follows:

Sensitivity - The sensitivity of the feature extraction and the feature classification is dictated by taking the proportion of a number of TP to the whole of TP and FN.

$$Sensitivity = \frac{TP}{(TP + FN)} \dots . (1)$$

Specificity- The specificity of the feature extraction and the feature classification can be assessed by taking the connection of a number of TN to the consolidated TN and the FP.

$$Specificity = \frac{TN}{(TN + FP)} \dots . (2)$$

Accuracy- The accuracy of feature extraction & feature classification can be figured by taking the proportion of true esteems shown in the populace.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \dots . (3)$$

Precision- Precision assesses what number of the data is classified to be Positive are really Positive by methods for the condition.

$$Precision = \frac{TP}{(TP + FP)} \dots . (4)$$

TP - Abnormal people are correctly recognized as Abnormal
TN - Normal people are correctly recognized as Normal
FP - Normal people are incorrectly recognized as Abnormal
FN - Abnormal people are incorrectly recognized as Normal

Individual classifiers performance parameters values are compared with hybrid machine learning classifier which is represented in below Table 1. Fig. 2 and Fig. 3 are shows the comparative performance analysis of Accuracy-Precision and specificity-sensitivity parameters respectively.

Table 1: PERFORMANCE OF INDIVIDUAL CLASSIFIERS WITH HYBRID LEARNING

Classifiers	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)
SVM	84	85	83	84
KNN	90	91	91	90
NB	91	92	91	90
Hybrid model (SVM+NB+KNN)	98	97	96	97

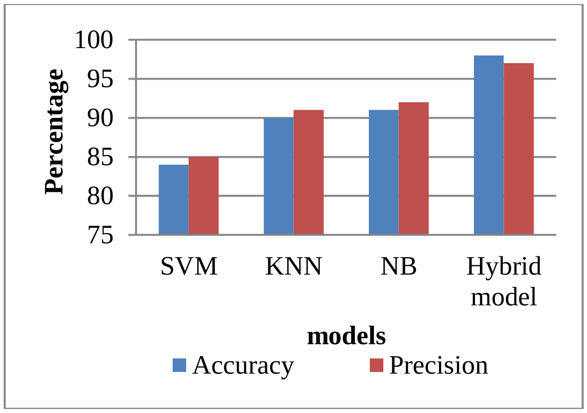

Fig. 2: COMPARATIVE ANALYSIS OF IN TERMS OF ACCURACY AND PRECISION PARAMETERS

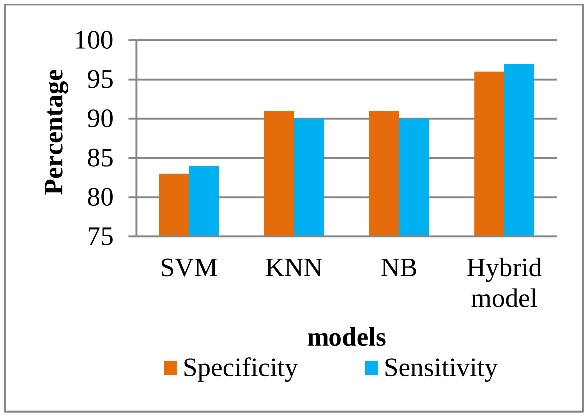

Fig. 3: COMPARATIVE ANALYSIS OF IN TERMS OF SENSITIVITY AND SPECIFICITY PARAMETERS

From results, the Hybrid Machine Learning model for diagnosis of CKD is efficient and shows high performance in terms of accuracy (98%), specificity (96%), sensitivity (97%), and precision (97%) than other models.

V. CONCLUSION

In this paper, Hybrid Machine Learning model for diagnosis of Chronic Kidney Disease (CKD) with optimal feature selection is described. Classification of kidney disease is vital for global improvement and accomplishment of practical guidance. For selecting ideal features think about ALO with better outcomes. The experiments are applied to the UCI Machine Learning Repository dataset. 75% of total dataset is used as training and remaining 25% of data is used for testing. Accuracy, Sensitivity, Precision, and specificity are used parameters. From results, the Hybrid Machine Learning model for diagnosis of CKD is efficient and shows high performance in terms of accuracy (98%), specificity (96%), sensitivity (97%), and precision (97%) than other models.

VI. REFERENCES

[1] Jing Wang, Xiao Wang, Yuanyuan Guo, Fei-Yue Wang, “A Parallel Medical Diagnostic and Treatment System for Chronic Diseases”, 2020 Chinese Automation Congress (CAC), Year: 2020

[2] N V Ganapathi Raju, K Prasanna Lakshmi, K. Gayathri Praharshitha, Chittampalli Likhitha, “Prediction of chronic kidney disease (CKD) using Data Science”, 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Year: 2019

[3] N. Asavakijthananont, M. Janyasupab, “Development of Nickel Nanowire on N-doped Carbon supported for Urea Measurement in Spent Dialysate for End-Stage Renal Disease Prognosis”, 2019 IEEE 19th International Conference on Nanotechnology (IEEE-NANO), Year: 2019

[4] Akash Maurya, Rahul Wable, Rasika Shinde, Sebin John, Rahul Jadhav, R Dakshayani, “Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning”, 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Year: 2019

[5] Muhsen, Heba, (2018). Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal, 3(2), 68-71.

[6] Widhiarso, Wijang, Yohannes Yohannes, and Cendy Prakarsah. (2018). Brain Tumor Classification Using Gray Level Co-occurrence Matrix and Convolutional Neural Network. IJEIS (Indonesian Journal of Electronics and Instrumentation Systems), 8(3), 179-190.

[7] Khawaldeh, Saed, (2017). Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. Applied Sciences, 8(1), 48-60

[8] Vijayarani, S., and S. Dhayanand. "Data mining classificationalgorithms for kidney disease prediction." International Journal onCybernetics & Informatics (IJCI) 4.4 (2015): 13-25.

[9] K.R Lakshmi, Y. Nagesh, and M. VeeraKrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering and Technology, vol.7, no.1, pp. 242-254, March 2014.

[10] G. Caocci, R. Baccoli, R. Littera, S. Orrù, C. Carcassi and G. La Nasa, "Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome", Artificial Neural Networks Kenji Suzuki, IntechOpen, DOI: 10.5772/53104, 2013