International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

An Effect of Machine Learning Based Classification Algorithms on Chronic Kidney Disease

Jerlin Rubini Lambert, Pramila Arulanthu, Eswaran Perumal

Abstract*: In the recent days, the prediction models of chronic kidney disease (CKD) becomes significant in the area of decision making which is helpful in healthcare systems. Because of large amount of medical data, efficient models are required to obtain precise results and data classification algorithms can be employed to detect the presence of CKD. Recently, various machine learning (ML) dependent on data classifier technique is presented for forecasting CKD. Since numerous classification algorithms for CKD prediction exist, there is a need to investigate the prediction performance of these algorithms. This paper propose a comparative analysis of 4 data classifier technique such as deep learning (DL), decision tree (DT), random forest (RF) and random tree (RT). The process of classification technique is analyzed with the help of reputed CKD dataset attained from UCI repository. From the simulation outcomes, it is evident that the DL method achieved optimal classifier action with respect to various namely accuracy, precision and recall.

Keywords* : Chronic Kidney Disease, Data classification, Deep Learning, Machine learning.

I. INTRODUCTION

Due to rapid increase in the biomedical and healthcare communities, a proper investigation of medical data is advantageous and leads to several advantages like earlier diagnosis, proper treatment and community services. A study conducted by McKinsey reported that half of the people in the US suffer from one or more chronic disease and almost 80% of US people spent money to take treatment on chronic diseases. Though living standard of the people increases, the chance of the existence of disease also increases. In recent days, chronic kidney disease (CKD) is considered as a main reason for the increase in the world's mortality rate. At the same time, it is not easy to identify the CKD quickly. To identify the existence of CKD at a faster rate, the medical data can be used to extract knowledge from the hidden structures.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Jerlin Rubini Lambert*, is Currently Pursuing Ph.D in Department of Computer Applications in Alagappa University, Karaikudi, Tamilnadu, India. E-mail: jel.jerlin@gmail.com.
Pramila Arulanthu is Currently Pursuing Ph.D in Department of Computer Applications in Alagappa University, Karaikudi, Tamilnadu, India. E-mail: pramimark@gmail.com.
Eswaran Perumal is Currently Assistant Professor in Department of Computer Applications in Alagappa University, Karaikudi, Tamilnadu, India. E-mail: eswaran@alagappauniversity.ac.in

Hence, it is needed to use computer-based systems to help doctors to predict diseases quickly [1].

The main issue in the diagnosis of diseases is the faster and precise prediction algorithm which requires less time to produce proper results. For the effective diagnosis results, it is more demanding to design a reliable and effective diagnosis model to support the rapidly increasing complicated diagnosis process. So, the computerized systems based on the classification algorithms needs the doctor's assistance to recognize the occurrence of disease easily.

Classifier is assumed to be a pattern which has been existed in the examination of data for extracting in a form of methods to explain the classes as well as to find the future information. Here, the classifying technique has a class label termed as exemplars or labels. The data mining process helps to determine a method with the application of labeled objects and applies them to detect the classes for unlabeled object. The main objective of a classifier is to detect the parameter which is named as categorical. The classifier has 2 stages as building the classifier and applies the classifiers in classifying operations. In general, classification model is categorized as 2 types such as supervised and unsupervised algorithms. The previous one has classification process which is carried out by employing the labeled classes. Furthermore, the training and testing data is modeled along with classes [2]. The alternate technique classifies data which is processed by applying the unlabeled classes. This method is used for including the class label with the current information. Also, clustering, association mining as well as dimensionality alleviations has been considered as data mining methods that belongs to unsupervised learning.

At present, various ML classifying technique has been presented in the research works [3]. In specific, the combination of diverse soft computing models assist to attain maximum outcome when compared with conventional models [4]. The Neural Networks (NN) is assumed to be well-known classifier that is obtained from domain of soft computing [5]. Also, NN is employed in several regions because of its simplicity, minimum processing expense as well as maximum computation. Multi-Layer Perceptron (MLP) is a form of feed forward NN, that helps to manage the classification issues present in a stochastic way.

Due to the availability of numerous classification algorithms for CKD prediction, there is a requirement to analyze the prediction performance of these algorithms. In this paper,

a comparison study is made among four data classification algorithms namely DL [6], DT, RF [7] and RT. The computation of classifying techniques has been examined under the application of standard CKD dataset which is obtained from UCI repository. It is clearly noted from the final outcome is that, DL technique has attained optimal classification methods with respect to diverse metrics like accuracy, precision and recall.

The remaining part of the paper is arranged as pursues: Section 2 explains the different classification algorithms utilized in the study. A detailed performance evaluation of these algorithms on CKD dataset is explained in Section 3. Section 4 and 5 explains the performance evaluation and results and discussion respectively. At last, the paper is concluded in Section 6.

II. LITERATURE SURVEY

Decision tree C4.5Classifier applied to 80 pregnant women information in the field of medical. Caesarian section had four categories and classified into two classes namely emergency and elective. The life of mother and baby had important during delivery time. It can be concluded with the mandatory factors of age, times of pregnancy, delivery time, heart condition as well as blood pressure. The delivery time could be divided into premature, timely and latecomer respectively. The Blood pressure has 3 constraints like low, normal, high. Hence, it attains a maximum of 86.25% accuracy including caesarian dataset [8].

Both attribute reduction and classification strategy used decision tree algorithm which applied to criminal behavior. This technique needed to take decision for police rapidly and precisely. The result of the hybrid algorithm had better than individual performance of the algorithms. Combination of decision tree and artificial neural network utilized for medical diagnosis. It included three components which are preprocessing and feature selection (MLP), knowledge production (DT), prediction with learned data for predicting human blood cell. Many medical researchers demonstrated that high risk factor of DNA viruses are strictly related to human cancers. The techniques of decision tree, Chi-squared Automatic Interaction Detection (CHAID) designed to construct a classification to forecasting breast cancer and fibro adenoma. Inexperienced physicians used medical classifier to prevent misdiagnosis.

DL enables processing models which is created from various processing layers to know the functions of information along with different levels of abstraction. These methods have obviously developed for the recent domains such as object detection, speech recognition, drug discovery and visual object recognition. Here, DL identifies difficult structure from massive datasets under the application of back propagation technique that denotes the process of machine exploitation from the inner accessories which is applied to measure the representation from all layers that has been existed in prior. From this paper a new DL technique have been obtained which is named as greedy deep weighted dictionary learning, which is mainly applied for diagnosis of medicinal data. Therefore, the presented weighted method is applied for correlating the sample and dictionary atom. Then, the demerit of over-fitting in patient classification of restricted training dataset is eliminated from conventional dictionary learning model with the help of 12-norm regularizing limitations. Also, DL is assumed to be an energetic technique that is organized from ML. It is capable of handling précised pattern analysis which is consumed from tedious original information. Both ML as well as DL is used to overcome the realistic issues under the application of NN. The variations among DL and ML could be identified as it consists of single hidden layer as well as one or more secret layers among input as well as output layers. Alternatively, classification and prediction of manipulated RF method is applied to detect the Unified Parkinson's Disease Rating Scale (UPDRS). Hence, the integration of parameters mined from microelectrode recordings (MERs) respectively [9].

III. MACHINE LEARNING BASED CLASSIFICATION ALGORITHMS

In this section, the four ML algorithms used for classification purposes such as DL, DT, RF and RT are discussed.

A. Deep Learning (DL)

DL is a ML method that enables the computer to teach in such a method that it is similar to humans, i.e. learn by examples [10]. It is the basic principle behind different interesting applications like automated cars, voice recognition in smart phones, TVs, etc. In recent days, DL become more popular and is used in medical institutions for the identification of the diseases at the previous stages. In DL, the automated scheme learns to do the classifier process straightaway from images, text, or sound. It leads to the achievement of better accuracy than the existing models and also sometimes it may exceed human performance. These models undergone training based on the labeled data and NN architectures which holds multiple layers. The basic structure of NN method is shown in Fig. 1. DL attains higher prediction performance when compared to other techniques and it assist users in different actual time applications.

While DL is introduced in the year of 1980, recently it became popular due to the following reasons:

- It needs massive amount of labeled data.
- It needs significant computing power. The high-performance GPUs with parallel architectures will be effective for DL. When DL is integrated to clusters or cloud computing, it enables the developers to considerably minimize the training time.

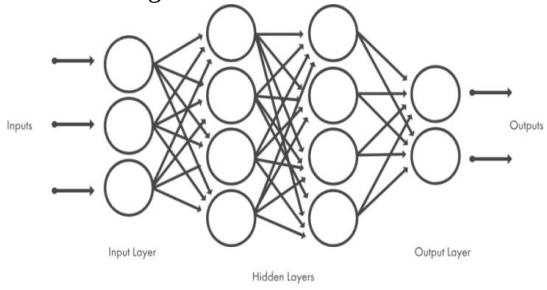

Fig. 1. NN structure

Since DL methods employ NN models, DL is mostly referred as Deep Neural Networks (DNN). The word deep signifies the entire number of secret layers from NN. The traditional NN technique has more than one hidden layer whereas the DN has more than 150 layers. DL is trained by a massive amount of labeled data and NN models which learns the features straightly from the data instead of manual feature extraction. This automated feature extraction behavior becomes more advantageous for DL and improves the overall accuracy. The overall process of the DL classification is shown in Fig. 2.

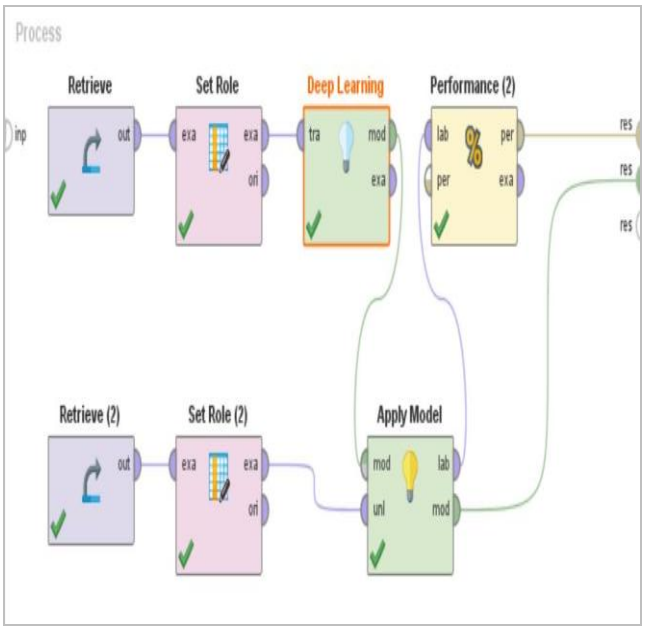

Fig. 2. Overall classification process of DL

B. Decision Tree (DT)

DT is said to be non-parametric supervised learning method which is applied for classification and regression process. The main intention is to generate a model for predicting the value of the intended variable using the simpler decision rules extracted from the data features. For example, in Fig. 3, the DT learns from the data to approximate a sine curve with a collection of if-then-else decision rules. When the depth of the tree increases, the complexity level of the decision rules will increase and the model becomes more fit.

DT is still popular due to the following characteristics: it is easier to understand and interpret. At the same time, the trees can be visualized and it needs only data less preparation time whereas the other methods needs data normalization process, creation of dummy variables and deletion of absent values. The cost of utilizing the tree is logarithmic in the number of data points employed to train the tree. It has the capability to manage numerical as well as categorical data. Since other methods are commonly specialized in the investigation of dataset involved with one variety of variables, DT manages multi-output problems.

The DT learning can generate over-complex trees which do not simplify the data and is termed as over fitting. The processes like pruning, placing lowest number of samples needed at a leaf node or fixing highest depth of the tree are needed to eliminate over fitting.

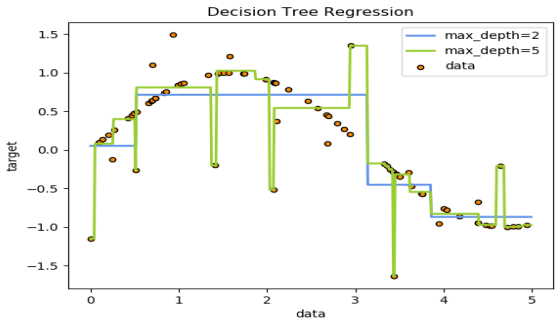

Fig. 3. DT model

C. Random Forest (RF)

The RF algorithm is a type of supervised classification algorithm [11]. As the name implies, RF algorithm generates the forest with more number of trees. The accuracy will be increased with more number of trees. It is an ensemble algorithm which indicates that these algorithms integrate more than one algorithms of similar or diverse type of object classifications. It generates a collection of DT from the randomly chosen subset of training set. It combines the votes from various DT to finalize the class of the test object. RFs grow with multiple classification trees. For the classification of new object from an input vector, the input vector is placed down at every tree in the forest. Every individual tree provides a classification and the forest selects the classification which holds more number of votes. Every individual tree will grow as follows:

- While many number of cases in training set is assumed to be N, the sample N cases present in arbitrary, whereas along with alternation, that is obtained from original data. The instance would be considered as training set in the process of developing a tree.
- If there is a existence of M input parameters, a number m<<M is represents in all nodes, m variables are selected in arbitrary manner from M as well as best split from the m has been executed to divide the node. Hence, value of m remains static while there is a growth in forest.

Every tree is grown to the largest extent probable with pruning process does not take place. There is no pruning.

The error rate of RF is based on the following two things:

- The association among other 2 trees in forest and forest error rate would be improved along with the increment of link between nodes.
- The strength of every tree in forest where the trees with minimum error would be known as optimal classifier. If there is a maximum improvement in strength of single tree lowers the forest error. Likewise, reduction in m lessens would result in association and the robust nature.

D. Random Tree (RT)

RT is developed by Leo Breiman and Adele Cutler [12], which deals with the classification as well as regression problems. RT is a group of tree predictors. The classification process of RT operates as follows: RT classifier gets the input feature vector, performs classification with each tree in the forest, and provides the class label that received the more number of "votes".

In case of regression, the classification response might have the maximum responses from entire trees present in the forest. Every individual tree is trained with the similar parameters but on diverse training set. The diverse sets are created from the raw training set by the use of the bootstrap process. For every training set, same number of vectors as in the raw set will be randomly selected (=N). The vectors are selected with replacement. This implies that some of the vectors will appear multiple times and some other vectors are not present. At every node of the trained tree, the parameters have not been applied for finding the optimal split, whereas it is referred to be random subset of them. Along with all nodes a novel subset has been produced. But, the size is allocated to every nodes as well as other tress. Then, it has been denoted as training parameter set for √ (number of variables) in a common factor. RT does not work in other any training computation. The RT has no necessity of precision measure such as cross-validation or a individual training set to determine the evaluation of training error.

IV. PERFORMANCE EVALUATION

A. Dataset Used

To investigate the classification performance of different ML algorithms like DL, DT, RF and RT, a benchmark own and original CKD dataset has been applied [13]. The definition and attribute information of dataset is provided in Table I and II respectively, where CKD dataset has 400 instances, 24 attributes as well as 2 classes. From 400 instances, 250 instances have been assigned along with existence of CKD and the rest of 150 instances have been modeled in lack of CKD.

Table- I: Dataset Information

Types of information	Description
Dataset	Chronic Kidney Disease
Source	Apollo Hospital
Number of instances	400
Number of attributes	24
Number of class	2
Class: CKD/NotCKD	250/150
Publisher	University of California Irvine Machine Learning Repository
Task	Classification.

Description of own and original CKD dataset are revealed in Table III. Age attribute is the age of patient .Kidney disease can affect any age of persons. Blood pressure has a standard range to find out how pulse is high or low. It has two values that contain systolic pressure and diastolic pressure. Here blood pressure considered as diastolic pressure only. Specific gravity provides information on the kidney's ability to concentrate in the urine. Albumin is a protein that found in the blood.

Damaged kidney passes albumin from blood to urine. Sugar is called glucose that present in the urine. High level of sugar passes from blood into urine, when the blood sugar is high. Red Blood Cells mean blood in urine. It may cause kidney infection or disease, urinary tract infection, kidney or bladder stones. Pus includes dead skin cells, bacteria and white blood cells. Pus cell and pus cell clumps are urinary tract infections (UTI). The tract may be kidney or urinary bladder. Bacteria are a sign of infection in the kidney. It moves from outside the body to the bladder.

Blood glucose random is used to check the level of glucose. It affects the people with diabetes. Blood Urea Nitrogen (BUN) test is referred to determine the amount of urea nitrogen in the blood. Serum creatinine, sodium and potassium are used to measure level of creatinine, sodium and potassium in blood respectively. Hemoglobin generally found in red blood cell count. Packed Cell Volume is a measurement of cells in blood. Blood is a combination of cells and plasma. As stated by American Association for clinical chemistry (AACC) Complete Blood Count (CBC) test measure the red blood cells, white blood cells and hemoglobin in the blood. White and red blood cell count issued to refer the number of white and red blood cells in blood respectively.

Table- II: Attributes Information of CKD Dataset

Medical Term	Attributes	Attribute Type and its value	Units
age	Age	Numerical	Years
bp	Blood Pressure	Numerical	mm/Hg
sg	Specific Gravity	Nominal (1.005,1.010,1.015, 1.020,1.025)	-
al	Albumin	Nominal (0,1,2,3,4,5)	-
su	Sugar	Nominal (0,1,2,3,4,5)	-
rbc	Red Blood Cells	Nominal (normal,abnormal)	-
pc	Pus Cell	Nominal (normal,abnormal)	-
pcc	Pus Cell Clumps	Nominal (present,notpresent)	-
ba	Bacteria	Nominal (present,notpresent)	-
bgr	Blood Glucose Random	Numerical	mgs/dl
bu	Blood Urea	Numerical	mgs/dl
sc	Serum Creatinine	Numerical	mgs/dl
sod	Sodium	Numerical	mEq/L
pot	Potassium	Numerical	mEq/L
hemo	Hemoglobin	Numerical	Gms
pcv	Packed Cell Volume	Numerical	%
wc	White Blood Cell Count	Numerical	cells/cumm
rc	Red Blood Cell Count	Numerical	millions/cmm
htn	Hypertension	Nominal(yes,no)	-
dm	Diabetes Mellitus	Nominal(yes,no)	-
cad	Coronary Artery Disease	Nominal(yes,no)	-
appet	Appetite	Nominal(good,poor)	-
pe	pedal edema	Nominal(yes,no)	-
ane	Anemia	Nominal(yes,no)	-

Blood pressure is interrelated with hypertension. Blood pressure that is high is called hypertension. Hypertension is considered to be a threat factor to several disease like heart attacks, stroke and kidney disease. Diabetes mellitus is the high level of blood sugar. Risk factors for coronary artery disease are high blood pressure, diabetes mellitus and kidney disease.

Hypertension, Diabetes mellitus, coronary artery and kidney disease are interconnected with one another. Decreased appetite may be known as poor appetite. Pedal edema is the build-up of fluid in the feet and lower legs. It causes the swollen of legs. Anemia is the lack of red blood cells or hemoglobin.

Test parameters of CKD dataset are shown in Table IV. Examination is known as test. Laboratory examination has blood and urine test. Laboratory examination is used to examine a sample of patient blood and urine. Blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count are blood test. Specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria are urine test.

Table- III: Description of Own and Original CKD Dataset

Name of the Attribute	Description of the Attribute
Age	Patients age is main criteria
Blood Pressure	The sign of heart rate should be measure
Specific Gravity	To measure the ratio of urine density compared with water density
Albumin	The total amount of protein in the urine is indicated.
Sugar	The high level sugar in the urine must show
Red Blood Cells	The high amount of red blood cells in urine must be specified
Pus Cell	Major and minor infections must be indicated
Pus Cell Clumps	Bunch of pus cells and infections to be identified
Bacteria	Identification of Kidney infection and growth level of bacteria
Blood Glucose Random	Glucose (sugar) level should be checked
Blood Urea	To measure the amount of urea nitrogen in blood
Serum Creatinine	To identify the amount of creatinine in blood
Sodium	To show the amount of sodium in blood
Potassium	To show the amount of potassium in blood
Hemoglobin	Protein in red blood cells must specify.
Packed Cell Volume	To measure the percentage of cells in blood
White Blood Cell Count	Finding the number of white blood cells in blood
Red blood cell count	Finding the number of red blood cells in blood
Hypertension	High level blood pressure must be indicated
Diabetes Mellitus	Must show the high level of blood sugar
Coronary Artery Disease	Heart disease must be identified which affects the kidney function
Appetite	Detecting the loss of appetite
Pedal Edema	Determination of legs swelling
Anaemia	Low level of red blood cells or haemoglobin must specify

Data types of attributes are shown in Table V. A numerical attribute is one that contains any value. (e.g age, blood pressure) A nominal attribute is one that contains only possible values (e.g albumin-0,1,2,3,4,5,bacteria-present, not present).Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count and red blood cell count are numerical type of attributes. Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anaemia are nominal type of attributes [14-16].

Table- IV: Test Parameters of CKD Dataset

Attributes	Type of Test
Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red blood cell count.	Blood
Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria.	Urine

Table- V: Datatypes of CKD Dataset

Attribute	Data types
Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count.	Numerical
Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema ,Anaemia	Nominal

Table- VI: Examination of CKD Dataset

Attribute	Types of Examination
Blood Pressure, Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, red blood cell count.	Laboratory
Hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, Anemia	Clinical

Examination of CKD Dataset is shown in Table VI. Examination is known as test. There are two types of test. It has laboratory and clinical examination. Laboratory examination is used to check a sample of patient blood and urine. It helps medical professional to diagnose medical conditions, evaluate treatment and monitor disease. Clinical examination is known as physical or medical examination. It contains the information about a person's medical history and symptoms experienced by the patient.

Flow chart of CKD dataset is shown in Fig.4.Every researcher finds a new problem for their research work. The problem of work may be medical, business, engineering, Industry etc. in any field. Researcher referred what are the disease is already present in repository apart from it and select new disease in medical field. Researcher got permission from higher officials, hospital administrative and hospital superintendent. Researcher met doctor/experts/specialist of that disease. Doctor taught basics knowledge, idea, symptoms, normal and abnormal value of the disease. Researcher referred multiple patient records which contain the value of the test report in medical record room. Researcher make table format and enter name and value of symptoms. It stored as row and column format in excel.

Each and every symptom has data type for dataset preparation. Both symptoms and attribute are same in medical and computer field.

Data type may be numeric or nominal. Some value of symptoms may be symbol and multiple values. So it converted into numeric or nominal. For example sugar: nil=0,trace=1,+=2 ,++=3, +++=4 ,++++=5 and rbc: val=1-2,2-4, 4-6,=normal, val= 6-8,8-10,10-12,plenty etc =abnormal are numeric and nominal type of conversion.

Data in excel format converted into arff format. ARFF is the attribute relation file format. Each statement of title, source, creator, guided by, date, number of instances and attributes, missing value, attribute and relevant information, class and class distribution are included before % symbol. Attribute, data type, measurement and values of nominal type are inserted in attribute information. Title of the dataset is inserted before @ symbol followed by relation keyword. For Example: @ relation chronic_ kidney_ disease. Each name of the attribute is added before@ symbol followed by attribute keyword. It is enclosed within single quotation with data type. For Example: @ attribute 'age' numeric. All value of the symptoms start with @ symbol followed by data keyword and values that separated by comma and end with class variable. For example: 48, 80, 1.020, 1, 0,?, normal, not present, not present, 121, 36, 1.2, ?, ?, 15.4, 44, 7800, 5.2, yes, yes, no, good, no, no, ckd. Sometimes patients cannot have the values of the symptoms or particular test cannot be examined. It considered as missing values that denoted by symbol.

Now the whole information of medical database prepared and run in the particular viewer (ex: arff viewer) without any error in WEKA machine learning software. Dataset submitted into donate dataset link in UCI machine learning repository.

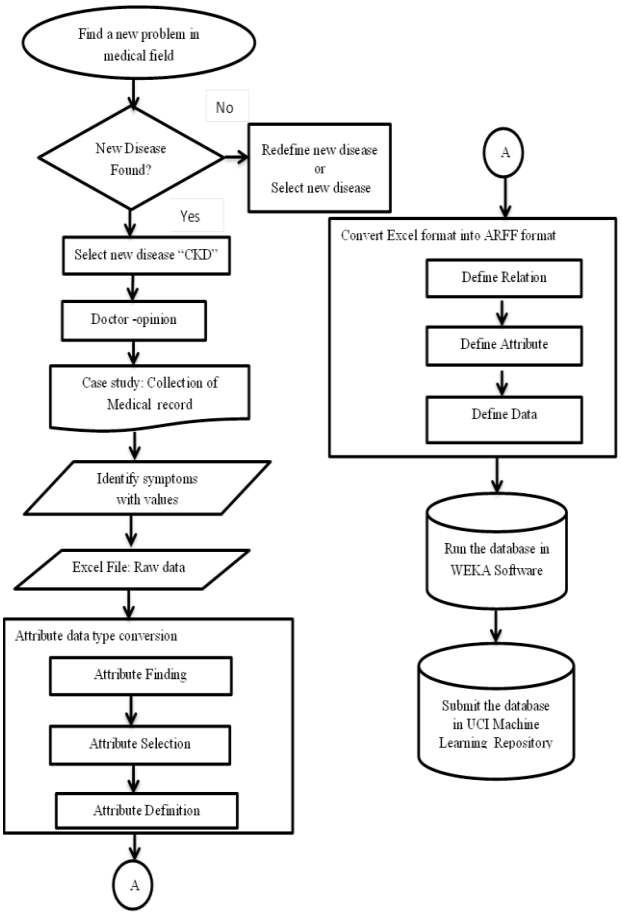

Fig. 4. Flow chart of CKD dataset

B. Metrics

Metrics are often applied to compare the attributes like accuracy, precision as well as recall. Before providing the definition of computation values, the method of confusion matrix has been tabulated in Table 7.

Table- VII: Confusion Matrix of CKD Prediction

CONFUSION MATRIX		
	Actual positive	**Actual negative**
Predicted positive	True positive(TP)	False positive(FP)
Predicted negative	False negative(FN)	True negative(TN)

Accuracy is defined as percentage of accurately classified instances. It also known to be vastly applied classifying process as well as the value must be nearer to 100 to accomplish optimal classification function.

$$\text{Overall predictive accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is defined as follows:

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

Mathematically, recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

V. RESULTS AND DISCUSSION

Table 8 explains the final outcome for classification process for various classifying models with respect to accuracy, precision and recall. Here, Fig. 5-7 denotes the relative outcomes of classifiers based on accuracy, precision and recall. In Fig. 5, it is proved that RT classifier reaches poor computation form alternate classification process. Simultaneously, RF obtains slightly better accuracy than RT. Similarly, DT handles to process quite-well to attain maximum accuracy when compared to RF and DT. Although DT performs well than alternate classification process, it unable to reach optimal outcome than DL. In overall, DL classification obtained the highest accuracy of 100 that is appreciably higher than the classifier accuracy achieved with other compared techniques.

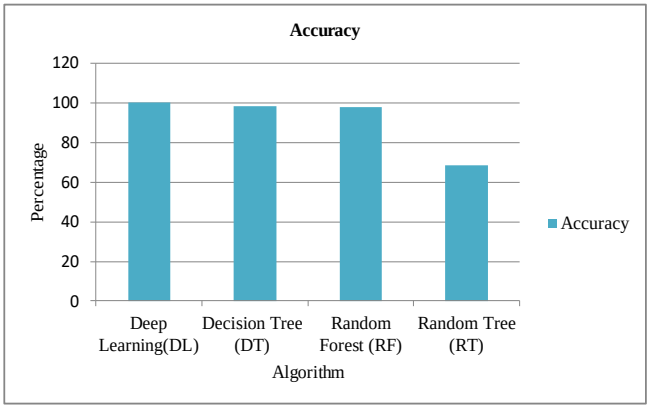

Fig. 5. Comparative analysis of different ML algorithms based on accuracy

The comparative outcomes of various ML techniques with respect to precision are shown in Fig. 6. From the figure, it is evident that the RT classification achieved the precision value of 54.48 that is much lesser than the other techniques. After that, DT and RF classifications revealed competitive action on one other. DT and RF classifications obtained precision value of 96.13 and 99.30 correspondingly. While RF controls to achieve optimal classifier outcomes than DT and RT, it fails to illustrate superior action over DL classification. The DL classification shows the highest precision value of 100. In overall, the DL classifications are found to be appreciably optimal than every compared techniques.

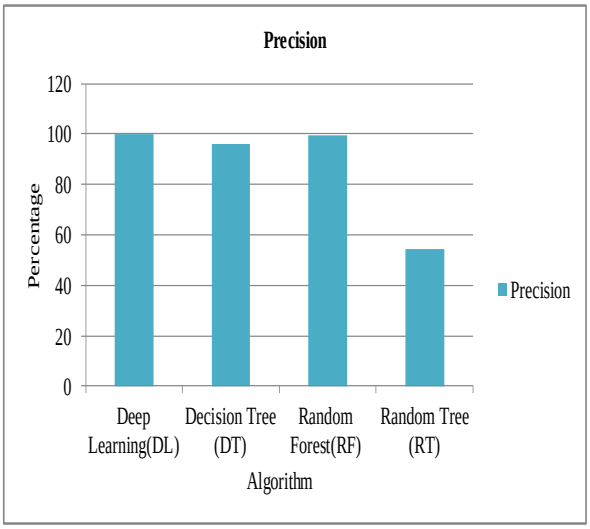

Fig. 6. Comparative investigation of different ML algorithms based on precision

Similarly, accuracy and precision has the relative analyses of different ML methods in terms of recall has

Table- VIII: Classification performance of different classifiers on CKD dataset

Algorithm	Accuracy	Precision	Recall
Deep Learning	100	100	100
Decision Tree (DT)	98.25	96.13	99.33
Random Forest (RF)	97.75	99.3	94.67
Random Tree (RT)	68.5	54.48	97.73

examined as well as relative outcome is depicted in Fig. 7. It is clearly noted from figure that it is significant that RF reached a lower value of 94.87 that is less than alternate techniques. Alternatively, RT exposes a considerable process than RF along with a recall measure of 97.73. However, DT performs in a better way such that RF as well as RT has the recall value of 99.33. In spite of DT tries to process in a standard manner than other techniques, it fails to produce optimal outcome than DL classification process. Hence, DL classifier reaches recall value of 100 that denotes a maximum classification process based on CKD dataset. The table and figures implies a clear statement that DL classifier accomplish a optimized classification operation than different other classification techniques. Therefore, it is concluded that DL classifier is optimal method for classifying instances of CKD dataset.

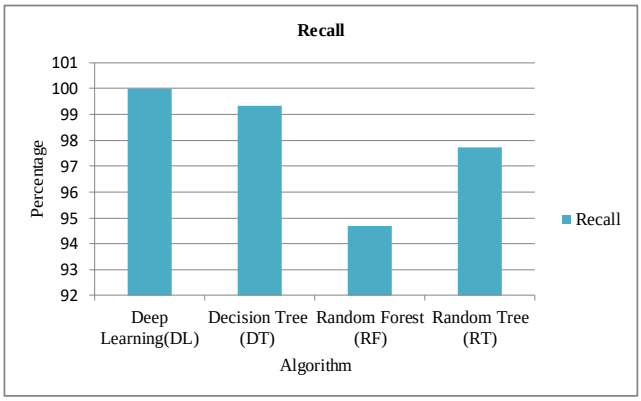

Fig. 7. Comparative analysis of different ML algorithms based on recall

VI. CONCLUSION

Presently, CKD is considered as a major reason for the increase in the world's mortality rate. At the same time, it is not easy to identify the CKD quickly. Recently, various ML dependent data classifier techniques are presented to forecast CKD. These papers propose a comparative analysis of 4 data classifier techniques such as DL, DT, RF and RT. The action of the classifier techniques are examined with the use of benchmark CKD dataset from UCI repository. From the simulation outcomes, it is evident that the DL method has achieved maximum classification with respect to accuracy, precision and recall. The table and figures denotes as clear statement that a DL technique represents the optimal classification measure when compared with alternate classification models. Hence, DL classification reaches a better classification accuracy, precision and recall around 100, 100 and 100 respectively. From entire process, it is finalized that DL classification has a better option for classifying the instances of CKD dataset. Therefore, the upcoming study would be expanded with the application of novel bio-inspired framework to classify CKD dataset.

ACKNOWLEDGMENT

This article has been written with financial support of RUSA–Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018.

We thank Dr.P.Soundarapandian, M.D., D.M (Senior Consultant Nephrologist) and Apollo Hospital, Karaikudi, India for their support and facility at the time of CKD data collection and preliminary study.

We thank UCI Machine Learning Repository for publishing attribute information of or own and original CKD dataset: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.

REFERENCES

1. Wu, R., Peters, W., Morgan, M.W.,"The Next Generation of Clinical Decision Support: Linking Evidence to Best Practice", *Journal of Healthcare Information Management*, vol.16,no.4,pp.50-55,2002.

2. P.J.Werbos,. Neuro control and Supervised learning: An overview and evaluation. In D.A.White & D.A *Sofge,editors, Handbook Intelligent Control :Neural ,fuzzy and adaptive approaches* Chapter3,pp.65-89,Van Nostrand Reinhold,Newyork, 1992.
3. Sudha, A., Gayathri, P., Jaisankar, N., "Effective analysis and predictive model of stroke disease using classification methods", *International Journal of Computer Applications*, vol.43, no.14,pp.26-31,2012.
4. Bashir, S., Qamar, U., Khan, F.H., Javed, M.Y,MV5: A clinical decision support framework for heart disease prediction using majority vote based classifier ensemble", *Arabian Journal for Science and Engineering*,vol.39,no.11,pp.7771-7783,2014.
5. Sneha Potghan,R. Rajamenakshi, Dr. Archana Bhise, "Multi-Layer Perceptron based Lung tumor Classification", *International conference on Electronics, Communication and Aerospace Technology (ICECA)*,pp.499-502,2018
6. Maruf, S., Javed, K. and Babri, H.A., 20"Improving text classification performance with random forests-based feature selection", *Arabian Journal for Science and Engineering*,vol.41, issue.3, pp.951-964,2016.
7. Farhad Soleimanian Gharehchopogh, Peyman Mohammadi, Parvin Hakimi," Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study", *International Journal of Computer Applications*, vol. 52, No. 6,pp.1-6, 2012.
8. Wang Hui, Wang Jing, Zheng Tao ,"Analysis of Decision Tree Classification Algorithm Based on Attribute Reduction and Application in Criminal Behavior",*IEEE*,pp.27-30,2011.
9. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Network., vol. 61, pp. 85–117, 2015.
10. Y. Le Cun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol.521, pp. 436-444,2015.
11. Khalilia, M., Chakraborty, S., Popescu M.,"Predicting disease risks from highly imbalanced data using random forest", *BMC Medical Informatics and Decision Making*, vol.11, no.51,pp.1-13,2011.
12. Cutler, A., Zhao, G., PERT-Perfect Random Tree Ensembles. *Computing Science and Statistics*, pp.1-8,2001
13. L.Jerlin Rubini and P.Eswaran, UCI Chronic Kidney Disease. *School of Information and computer Sciences*, University of California, Irvine, CA, USA. (2015).Chronic kidney disease dataset Link :https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.
14. L.Jerlin Rubini and P.Eswaran, "Generating comparative analysis of early stage prediction of chronic kidney disease", *International Journal of Modern Engineering Research*, vol.50, pp.49-55, July 2015.
15. L.Jerlin Rubini and P.Eswaran, "Comparative Analysis of Decision Tress classifiers for Machine Learning of chronic kidney disease", *International Journal of Applied Engineering Research*, vol.10, no.82, pp.570-576, 2015.
16. L.Jerlin Rubini, P.Eswaran ,"Optimal fuzzy min-max neural network for medical data classification using group search optimiser algorithm",*International Journal of Mobile Network Design and Innovation*, vol. 7, Nos. 3/4, pp.140-149,2017

AUTHORS PROFILE

Jerlin Rubini Lambert received her B.Sc degree from the Mother Teresa University, Kodaikanal, Tamil Nadu, India in 2007, MCA degree from the Anna University, Chennai, Tamil Nadu, India in 2010, and MPhil degree from the Thiruvalluvar University, Vellore, Tamil Nadu, India in 2011. From 2011 to 2012, she worked as a Lecturer. Currently, she is PhD part time Research Scholar in the Department of Computer Science and Engineering, Alagappa University and an Assistant Professor in the Department of Information Technology in Puducherry. His research interests are in data mining, software engineering and networks.

Pramila Arulanthu received her B.Sc degree from Idhaya College for Women, Sarugani in 2013, B.Ed. Degree from Sri Raaja Raajan College of Education for women in 2014, and M.Sc. degree from Alagappa University from 2014-2016,and M.Phil. degree from Alagappa University in 2017.Currently , she is the PhD full time Research Scholar in the Department of Computer Applications, Alagappa University. Her research interests are in data mining, image processing, and networks.

Eswaran Perumal received his M.Sc in Computer Science and Information Technology from the Madurai Kamaraj University, India in 2003, M.Tech and PhD in Computer and Information Technology from the Manonmaniam Sundaranar University, India in 2005 and 2010,respectively. In 2010, he joined the Department of Computer Science and Engineering, PSN College of Engineering and Technology, as an Assistant Professor. Since May 2012, he has been with the Department of Computer Science and Engineering, Alagappa University, Karaikudi, India, as an Assistant Professor. He was the recipient of the Junior Research Fellowship Award of the University Grants Commission, New Delhi in 2008. He has published more than 25 scientific papers in the field of digital image processing and data mining. His research interests include digital image processing, focusing on color image edge detection, data mining and computer vision.