MACHINE LEARNING ALGORITHM SELECTION FOR CHRONIC KIDNEY DISEASE DIAGNOSIS AND CLASSIFICATION

M.Gokiladevi [1*], Sundar Santhoshkumar [2], Vijayakumar Varadarajan[3]

[1]Doctoral Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, 630 003, India

[2]Assistant Professor, Department of Computer Science, Alagappa University, Karaikudi, 630 003, India

[3]School of Computing Science and Engineering, The University of New South Wales, Sydney, Australia

Email: kokilamurugesan80@gmail.com[1*] (corresponding author), santhoshkumars@alagappauniversity.ac.in[2] vijayakumar.varadarajan@gmail.com[3]

DOI: https://doi.org/10.22452/mjcs.sp2022no1.8

ABSTRACT

In last decades, chronic kidney disease (CKD) becomes a global health problem that is steadily developing worldwide. It is a chronic illness highly related to increased morbidity and mortality, cardiovascular diseases, and high healthcare cost. Earlier identification and classification of CKD is treated as a major factor in controlling the mortality rate. Data mining (DM) techniques are used for the extraction of hidden details from the clinical and laboratory patient data that is used to aid doctors in enhancing diagnostic accuracy. Recently, machine learning (ML) techniques are commonly employed for the prediction and classification of diseases in healthcare sector. With this motivation, this study examines the performance of different ML algorithms to diagnose CKD at the earlier stages. The proposed model involves data pre-processing in two stages such as missing value replacement and data transformation. Besides, a set of five ML based classification models are involved such as support vector machine (SVM), random forest (RF), logistic regression (LR), K-nearest neighbor (KNN), and decision tree (DT). For investigating the performance of the different ML models, a benchmark CKD dataset from UCI repository is employed and the results are examined under different aspects. Among the different classifiers, the RF model has accomplished superior results with the maximum precision of 0.99, recall of 0.99, and F-score of 0.99 with a minimal error rate of 0.012.

Keywords: ***Data Mining, Data Classification, Chronic Kidney Disease, Machine learning, Data* Pre-*processing, Healthcare***

1.0 INTRODUCTION

Recent developments in information technology (IT) including wearable computing, big data, mobile communication system, and Internet of Things (IoT), are widely utilized from healthcare field to improve healthcare services [1]. Similarly, the exponential growth in medicinal information results in various problems to manage, store and process the data. Nowadays, lower grade inflammation is processed as a significant feature of chronic kidney disease (CKD). A patient with CKD is high possibility of affecting heart disease [2]. The early phase of CKD doesn't display some significant symptoms and highly complex for detecting without analysis such as blood and urine tests. If the CKD is identified at the early stage, better treatment and preventive actions can be provided for controlling the possibilities of dialysis/transplantation. A research stated that the early recognition of CKD can decrease the development of disease with nurses from expert of nephrology and healthcare physicians. In general, imaging methods are utilized for identifying the existence of CKD. However, due to a huge number of persons, it isn't possible to test every patient, and person with a high chance of having CKD would be suggested to endure broad testing. Presently, the maintenance of medical databases turned into a complex task in the field of healthcare [3]. The person's data hold district features and diagnoses relevant to disease should be providing with great significance for attaining higher quality of service. As the data documented in the clinic database might have lost and unwanted data, it turns into difficult for mining personal data.

Determining the disease stages is critical since it provides various signs which assist the establishment of needed treatment and interventions [4, 5]. Thus, data mining could perform a significant part in removing hidden data in huge person clinical and medical dataset that doctors often collects from patient to attain perceptions regarding the diagnosis data, and for implementing accurate treatment strategies. The prediction characteristic denotes the processing method of a trained set having a group of outcomes and attributes. The classification process is commonly a supervised learning procedure which assumes a link among class labels and features [6]. A classification and prediction method uses the trained data for creating a module and test data to examine the predictive efficiency.

In recent times, artificial intelligence (AI) methods are utilized for improving the presented classification module. ML methods have been utilized for predicting and classifying in the field of healthcare. [7] utilized SVM method for classifying predicting diabetes and pre-diabetes persons, and the outcome shows that SVM is beneficial to categorize persons with general diseases. Likewise, [8] have categorized Alzheimer's disease by SVM for analyzing entire brain anatomical magnetic resonance imaging (MRI) for several persons, and the result displays that SVM is a significant method for Alzheimer's disease earlier recognition. [9] have made heart disease prediction by the NB, NN, and DT technique, and PRNN gives an optimum outcome related to another technique for predicting heart disease. [10] proposed a predictive of HBV induced liver cirrhosis utilizing Multilayered Perceptron (MLP) method and the outcomes demonstrates that the MLP classification provides suitable prediction output for liver disease, mainly in HBV relevant liver cirrhosis persons.

2.0 LITERATURE REVIEW

This section performs a detailed review of state of art ML models for CKD diagnosis and classification. Qin et al. [11] proposed an ML method to diagnose CKD. The CKD dataset was attained from the UCI repository that has a huge number of missing values. The KNN imputation was utilized for filling the missing values that chooses numerous complete instances with equivalent measurement for processing missing data for every incomplete instance. Missing values are commonly viewed in real life medical situations since person might miss few measurements for several reasons. When efficiently complete the incomplete dataset, 6 ML methods like RF, LR, KNN, SVM, FFNN, and NB have been utilized for establishing the modules. To rapidly forecast the seriousness of CKD utilizing available demographic and blood biochemical features in follow-ups, Xiao et al. [12] related and established various prediction modules by statistical, ML and NN methods comprising KNN, Elastic Net, RF, lasso regression, SVM, ridge regression, NN, XGBoost and LR.

Sobrinho et al. [13] investigated the use of ML methods for assisting earlier diagnoses of CKD in emerging countries. Quantitative and qualitative relative analyses are, correspondingly, accompanied by a systematic study and research with ML methods, with k fold cross validation (CV) technique. This analysis allows a discussion on the appropriateness of ML methods to screen CKD threat, concentrating on lower income and hard for reaching setting of emerging countries, because of the particular issues confronted by them, for example, insufficient primary healthcare. Jongbo et al. [14] utilized 2 ensembles methods as Bagging and Random Subspace approaches on 3 base learners NB, DT, and KNN to improve the classification efficiency of the modules. Preceding classification, data pre-processing has been executed for handling data scaling and missing values to normalize the range of autonomous parameters.

Segal et al. [15] examined 10,000,000 medical insurance claims from 550,000 person recordings by a commercial health insurance database for CKD diagnoses. It utilized a feature embedding technique depending upon the execution of the Word2Vec method for capturing additional temporal data to the 3 major modules of the data: medications, diagnosis, and processes. For the analyses, they utilized the gradient boosting tree method (XGBoost execution). Wang et al. [16] analyzed the CKD by ML methods according to the dataset from UCI ML data warehouse. In several classification methods comprising J48, ZeroR, IBk (KNN), NB, and OneR. The dataset is pre-processed via normalizing and completing missing data.

In Ma et al. [17], Heterogeneous Modified Artificial Neural Network (HMANN) was presented to earlier diagnosis, detection, and segmentation of chronic renal failure on Internet of Medical Things (IoMT) architecture. Additionally, the presented HMANN is categorized as an SVM and MLP with BP technique. Qin et al. [18] proposed an ML method to diagnose CKD. The KNN imputation was utilized for filling in the missing values that choose various complete instances with the equivalent measurement for processing the missing data to all incomplete instances. When efficiently completing the incomplete dataset, 6 ML methods like FFNN RF, NB, KNN, SVM, and LR have been utilized for establishing the modules. Amongst these ML modules, RF attained an optimum efficiency with 99.75% diagnosis accurateness. Elhoseny et al. [19] established an intellectual forecast and classifier

system for healthcare, such as Density based Feature Selection (DFS) with Ant Colony based Optimization (D-ACO) method for CKD.

3.0 DIAGNOSIS MODELS

In this section, the five different ML models are discussed in detail. Fig. 1 depicts the overall working of the proposed model. The diagnostic process involves two major stages such as data pre-processing and classification. Besides, a set of five ML based classification models are involved such as support vector machine (SVM), random forest (RF), logistic regression (LR), K-nearest neighbor (KNN), and decision tree (DT). These ML models are commonly available for classification process and therefore they are used in this study. The detailed working of these processes is offered in the succeeding sections.

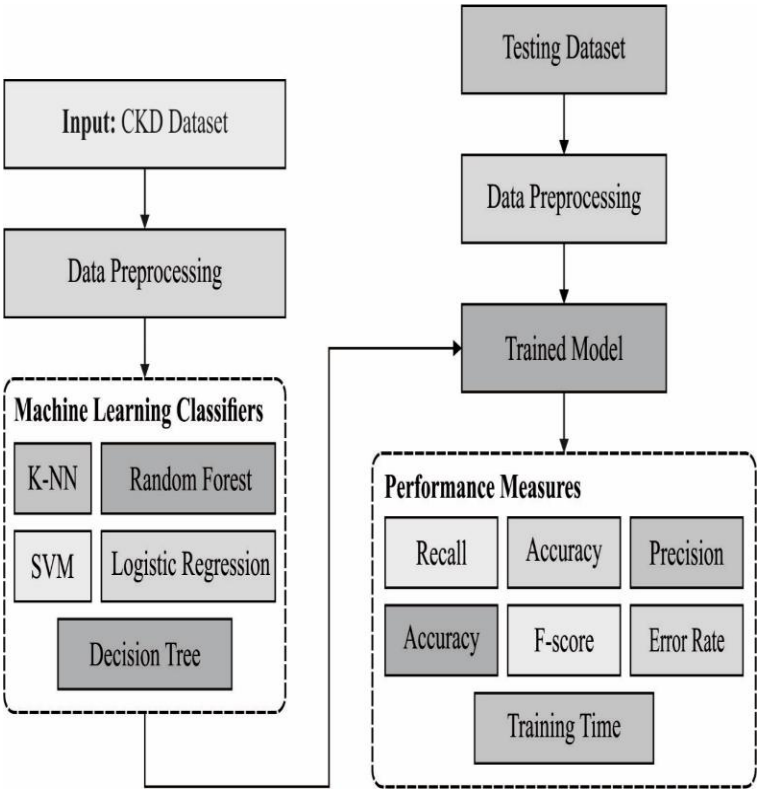

Fig. 1: Working process of proposed model

3.1 Data Pre-processing

The data pre-processing involves missing value replacement and data transformation. Firstly, median technique is used for the imputation of the missing values. In this technique, the missing values are filled with the median value of the whole feature columns. If the data is skewed, it is beneficial to use median values for the replacement of the missing values. Then, data transformation process takes place where the numerical values that exist in the dataset are transformed into categorical values (0 and 1 or YES and NO). The pre-processed data are fed into the ML models to detect and classify the presence of CKD.

3.2 SVM Model

In recent years, the SVM is the most extensively utilized data learning method. It is generally utilized for addressing binary pattern classification problems. The binary SVM creates a group of hyper plane in an infinite dimensional space that could be separated into 2 types of representations, like linear and non-linear SVM. Initially, they deliberate a binary classification problem; the trained dataset $T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_l, y_l)\}$, $y_i \in \{-1, 1\}$,

$x_i \in R^d$, whereas x_i denotes data point and equivalent y_i represents designed label. The l indicates amount of components in the trained dataset. Fig. 2 illustrates the optimal hyperplane of SVM. The linear SVM detects an optimum separate margin by resolving the succeeding optimization process:

$$Minimize\left\{\frac{1}{2}|w|^2 + C\sum_{i=1}^{l}\varepsilon_i\right\}, \varepsilon_i \geq 0 \qquad (1)$$

$$Subject\ to\ y_i(w^T x_i + b) \geq 1 - \varepsilon_i, i = 1, 2, \cdots, l, \qquad (2)$$

Where C denotes penalty value, $\varepsilon_?$ indicates positive slack parameters, w represents normal vector, and b indicates scalar quantity. The minimal problem is decreased by Lagrangian multiplierα_i that could attain an optimal based on Karush-Kuhn-Tucker state [20]. When$\alpha_i > 0$, then the equivalent data x_i so-called support vector (SV), and, thus, the linear discriminate function is stated with the optimum hyperplane variables w and b in the succeeding formula:

$$f(x) = sgn\left(\sum_{i=1}^{l}\alpha_i\, y_i x_i^T + b\right). \qquad (3)$$

Eq. (1) is converted to (4) by its unrestricted dual formula:

$$Maximize\ \left\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\,\alpha_j y_i y_j x_i x_j\right\}, \qquad (4)$$

$$C \geq \alpha_i \geq 0, i = 1, \ldots, l, \sum_{i=1}^{l}\alpha_i\, y_i = 0. \qquad (5)$$

Eq. (4) could be resolved but the quadratic programming methods and stationary Karush-Kuhn-Tucker state. The resultant solution W denotes linear integration of the trained vector and b represents average of whole SV is given by

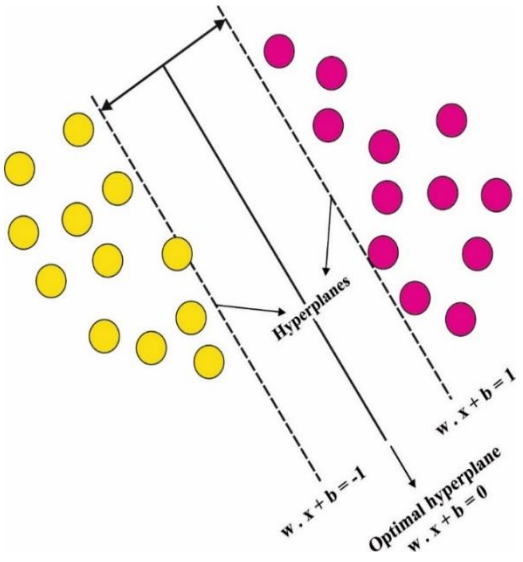

Fig. 2: SVM Optimal hyper plane

$$w = \sum_{i=1}^{l}\alpha_i y_i x_i, \qquad (6)$$

$$\mathrm{b} = \frac{1}{N_{\mathrm{SV}}} \sum_{i=1}^{N_{\mathrm{SV}}} (W\mathrm{x}_i - y_i) ,$$

Where N_{SV} denotes amount of SV. The linear SVM is extended to the non-linear cases by substituting $\mathrm{x_i}$ with mapping to the feature space $\theta(\mathrm{x_i})$, simultaneously, the $x_i^T \mathrm{x}$ denotes form of $\theta(\mathrm{x_i})^T \theta(\mathrm{x_i})$ in the feature space. Therefore the non-linear discriminate function is given by:

$$f(\mathrm{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right); \qquad (7)$$

Where $K(x_i, x) = \langle \theta(x_i), \theta(x) \rangle$ and $K(\mathrm{x_i}, \mathrm{x})$ indicates kernel function. The extensively utilized kernel function is

the radial basis function (RBF), as a result of its reliable and accurate efficiency that is given by

$$K(\mathrm{x_i}, \mathrm{x}) = \exp\left(-\gamma \|x_i - x\|\right)^2. \qquad (8)$$

The γ indicates predefined smoothness variable which controls the width of RBF kernel; so, (4) is given but

$$Maximize \left\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \, \alpha_j y_i y_j \exp\left(-\gamma \|x_i - x\|^2\right) \right\},$$

$$C \geq \alpha_{\mathrm{i}} \geq 0, i = 1, \ldots, l, \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} \, y_{\mathrm{i}} = 0. \qquad (9)$$

i. Non-peer-reviewed research such as tutorials, reports, lectures, and editorial articles**.**

ii. Studies with participating adolescents and children.

iii. Studies are based on conceptual frameworks or structures without analytical or empirical results.

Two researchers compared their results, and when there was a discrepancy between two results from the same paper, a joint review was performed with all researchers and an agreement was reached to include or excluded the study from the systematic review. Finally, 39 studies were excluded and a total of 57 primary studies were identified for inclusion in this systematic review.

3.3 RF Classifier

RF is an integration of tree predictors thus all trees are dependent upon the values of arbitrary vector experimented individually and similar distribution for every tree in the forest. The RF is an ensemble classification that denotes a novel method which utilizes several classifiers. The idea is that integrating ensemble classification is frequently more precise compared to other ensembles [21], avoid the conflicts between feature subsets. Consequently, the RF classifiers are extensively utilized in process remote sensor images. The general component in entire processes is for k-th tree, an arbitrary vector θ_k is created independent of previous arbitrary vectors $\theta_1, \ldots, \theta_{k-1}$, however with similar distribution; the tree is developed by trained set and θ_k, resultant in a classification $h(x, \theta_k)$, whereas x denotes input vector. The classification contains many trees created analytically by pseudo random selection subset of elements of the feature vector, specifically, trees are made in arbitrarily selected subspaces preserving the maximum accurateness of trained data and enhancing generalization accurateness as it develops in difficulty. Over a bagging procedure, if the RF classifier creates a tree develop, it utilizes an optimum split of an arbitrary subset of input feature/prediction parameter in the separation of all nodes, rather than utilizing optimum split parameters.

Thus, weakening the strength of all single trees decreases the connection among trees and generalization errors. Because the trees of RF classification grow without pruning, the time to generate a module doesn't significantly increase. The predictive class of observation is estimated depending upon most vote of trees from RF module, and discrimination function is given in Eq. (10):

$$H(x) = argmax_{\Upsilon} \sum_{i=i}^{k} I(h_i(X, \theta_k) = \Upsilon) \qquad (10)$$

Where $I(\cdot)$ represents indicator function, $h(\cdot)$ denotes single DT, and Y indicates output parameter, $argmax_{\Upsilon}$ indicates Υ value if increasing$\sum_{i=i}^{k} I(h_i(X, \theta_k) = \Upsilon)$. Every novel trained set is created for making a tree grow, one third of the instance are excepted arbitrarily, named out of bag (OOB) instance. The residual (in the bag) instances are utilized to build the tree. These OOB instances are utilized for evaluating the module efficiency and verified that the OOB evaluations are unbiased.

3.4 LR Classifier

The LR is utilized if there is a need for predicting likelihoods of absence/presence of a certain characteristic, disease, or result in common depending upon a group of autonomous explanatory parameters of various kinds like categorical continuous, or discrete [22]. As the predictive likelihood should lie in-between zero and one, modest linear regression methods are inadequate for attaining that, as they permit the dependent parameter for passing this limit and generate unreliable outcomes. Determined as P_1, the likelihood of an object belongs to group one, and P_0, the likelihood of an object belongs to group zero. It is given by

$$z_i = \log\left(\frac{P_{i1}}{P_{i0}}\right) \qquad (11)$$

$$= b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik},$$

Where P_{i1}/P_{i0} denotes odds ratio, b_j indicates value of jth coefficient, $j = 1, \ldots, k$, and x_{ij} represent value of ith case of jth predictor. The variables (b_o to b_k) of the logistic module is calculates by highest probability technique. The likelihood of events to happen is calculated by

$$P(Y_i = 1 \setminus X_i) = \frac{e^{b^T X;}}{1 + \left(e^{b^T X_i}\right)} = \frac{1}{1 + e^{-b^T X_i}}, \qquad (12)$$

Where $e^{b^T X_i}$ denotes linear predictor of the LR function, and Y_i indicate event in study (dependent parameter). When they utilize a likelihood cutoff 0.5, then they could categorize an object to group one when the calculated $P_1 > 5$ and to group zero when $P_1 < 5$. To calculate the variables of LR method the highest probability increases the coefficient of the log probability function, a statistic that summarizes the data of the predictor parameter. Both logistic and linear discriminant regression analyses take a similar functional frame; a combination of independent parameters and a rule for classification. However, it contains several variances regarding assumptions created in employing them in dataset. Concerning discriminant analyses, the assumption has great comparison with the assumption created for regular regression and (i) independent parameter should contain multivariate normal distribution, therefore it allows continuous/ratio parameters for entering the analyses and excludes every formation of categorical parameter, (ii) the variance and covariance matrix of every independent parameter should be homogenic amongst the population group separated with dependent parameter, and (iii) independence of the case.

3.5 KNN Classifier

KNN classifier is a simple non-parametric technique for classification. In spite of the easiness of this technique, it executes very well, and significant model. KNN classification needs a metric d and positive integer K. KNN rules

hold the location of trained instances and their class. If the decisions regarding novel incoming data are required, distance among trained instances and query data are estimated. Depending upon determined threshold for the rule (K number), K instance with minimum distance is chosen and the class with extra instances inbound is the outcome. Simultaneously, for instance, when there are 2 or 3 features to classifier condition, location of trained and input samples are visualized on two and three dimensions Cartesian coordinate [23]. KNN is a classification and its accurateness is often 100% on trained dataset, as the location of trained instances and their class are constant in the classifiers procedure. In this study, parameter K value is utilized among one and twenty.
In KNN classification, Euclidean distance metric is a simple and easy for implementing technique to compute distance in multi-dimension input space that could produce modest results than sophisticated ML techniques. The Euclidean distance among points p and q is the length of line among themselves. In Cartesian coordinate, when p_i and q_i are 2 points in Euclidean n-space, next the distance from p to q as follows

$$d_{E=}\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (13)$$

3.6 DT Classifier

The DT are identified between the most efficient and effective ML methods and effectively employed for resolving real-time challenges with AI fields. A plethora of methods was presented to make DT from provided trained set and guarantee the classification of query samples [24]. The most utilized method follows a top-down Induction of Decision Tree approach (TDIDT) which contains a recursive divide and conquer approach by succeeding stage:

- Choose attribute selection measure, the attribute allows an optimum probable dividing of the trained set;
- Divide the present trained data into trained subset based on selected attribute values.
- Suggest a trained subset as leaves while an end condition is attained.

Regarding the attribute selection procedure, numerous measures are presented. The data gaining, measure the effectiveness of an attribute while categorizing the trained instance, is one between the optimum known and extensively utilized measure. Assumed a trained data S and attribute A, the data gain is given by:

$$Gain(S, A) = Info(S) - Info_A(S) \qquad (14)$$

Where

$$Info(S) = -\sum_{i=1}^{Q} p_i \cdot log_2 p_i \qquad (15)$$

and

$$Info_A(S) = -\sum_{v \in Domain(A)} \frac{|S_v^A|}{|S|} \qquad (16)$$

Where p_i reflect the proportion of objects containing θ_i as class (i.e. $i \in \{1, \dots, Q\}$) and S_v^A correspond to the trained subset where the attribute A has v as value.
The main constraint of this measure is that the attribute with maximum values is the most stimulated one. This led to the summary of the $GainRatio$ measures utilized in the C4.5 method. It can be represented by:

$$GainRatio(S, A) = \frac{Gain(S, A)}{\text{SplitInfo(A)}} \qquad (17)$$

Where

$$SplitInfo(A) = \sum_{v \in Domain(A)} \frac{|S_v^A|}{|S|} . log_2 \frac{|S_v^A|}{|S|}. \qquad (18)$$

The search for this review was carried out from the earliest studies to December 2020. This involves defining all the nonclinical techniques for the diagnosis of MetS. Based on the distribution in Figure 2, the number of searches for nonclinical methods peaked in 201 . The trend of interest in research has decreased by 10% from 2016 to present. Two studies were published in 2020, confirming the inclusion of up-to-date related studies in this systematic review.

4.0 PERFORMANCE VALIDATION

This section examines the performance of the ML models on the applied benchmark CKD dataset from UCI repository. The details of the CKD dataset are provided in Table 1. The dataset encompasses a set of 400 instances with 24 features and 2 classes. Among the 400 instances, 62.50% of the instances fall into positive class and the remaining 37.50% of the instances come under negative class. Besides, the description of the attributes involved in the dataset is given in Fig. 3. In addition, the heat map regarding the features that exist in the dataset is depicted in Fig. 4.

Table 1: Dataset Description

Description	**Values**
Num. of Samples	400
Num. of Features	24
Num. of Class	2
Percentage of Positive instances	62.50%
Percentage of Negative instances	37.50%
Data source	UCI

S. No	Features	Datatype	Features Description	Unit of Measurement
1	Age	Numerical	Age	Years
2	bp	Numerical	Blood Pressure	mm/Hg
3	sg	Nominal	Specific Gravity	1.005, 1.010, 1.015, 1.020, 1.025
4	al	Nominal	Albumin	0, 1, 2, 3, 4, 5
5	su	Nominal	Sugar	0, 1, 2, 3, 4, 5
6	rbc	Nominal	Red Blood Cells	Normal, Abnormal
7	pc	Nominal	Pus Cell	Normal, Abnormal
8	pcc	Nominal	Pus Cell clumps	Present, Not_Present
9	ba	Nominal	Bacteria	Present, Not_Present
10	bgr	Numerical	Blood Glucose Random	mgs/dl
11	bu	Numerical	Blood Urea	mgs/dl
12	sc	Numerical	Serum Creatinine	mgs/dl
13	sod	Numerical	Sodium	mEq/L
14	pot	Numerical	Potassium	mEq/L
15	hemo	Numerical	Haemoglobin	gms
16	pcv	Numerical	Packed Cell Volume	0,1, 2…
17	wbcc	Numerical	White Blood Cell Count	cells/cumm
18	rbcc	Numerical	Red Blood Cell Count	millions/cumm
19	htn	Nominal	Hypertension	Yes, No
20	dm	Nominal	Diabetes Mellitus	Yes, No
21	cad	Nominal	Coronary Artery Disease	Yes, No
22	appet	Nominal	Appetite	Good, Poor
23	pe	Nominal	Pedal Edema	Yes, No
24	ane	Nominal	Anemia	Yes, No
25	Class	Nominal	CKD, Not_CKD	CKD, Not_CKD

Fig. 3: Attributes Descriptions

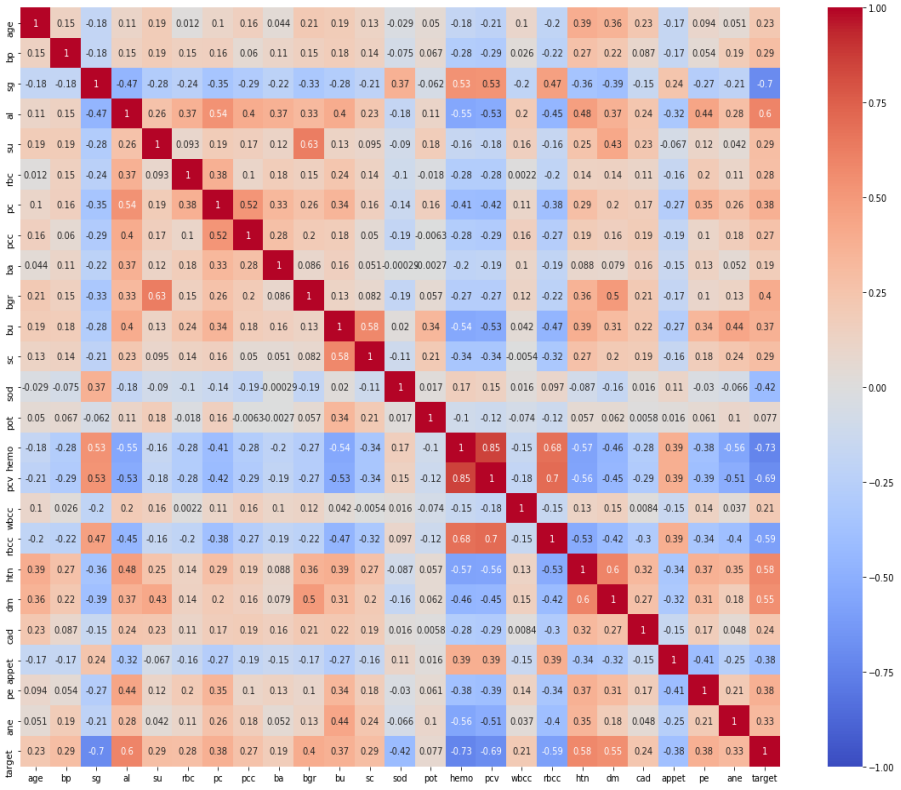

Fig. 4: Heat Map of the CKD Dataset

Fig. 5 depicts the confusion matrices generated by different ML models on the CKD dataset. Fig. 5a shows that the RF model has identified a set of 28 instances as negative and 51 instances as positive. In addition, Fig. 5b shows that the KNN model has identified a set of 26 instances as negative and 28 instances as positive. Moreover, Fig. 5c

shows that the DT model has identified a set of 25 instances as negative and 52 instances as positive. Furthermore, Fig. 5d shows that the SVM model has identified a set of 28 instances as negative and 31 instances as positive. At last, Fig. 5e shows that the LR model has identified a set of 28 instances as negative and 49 instances as positive.

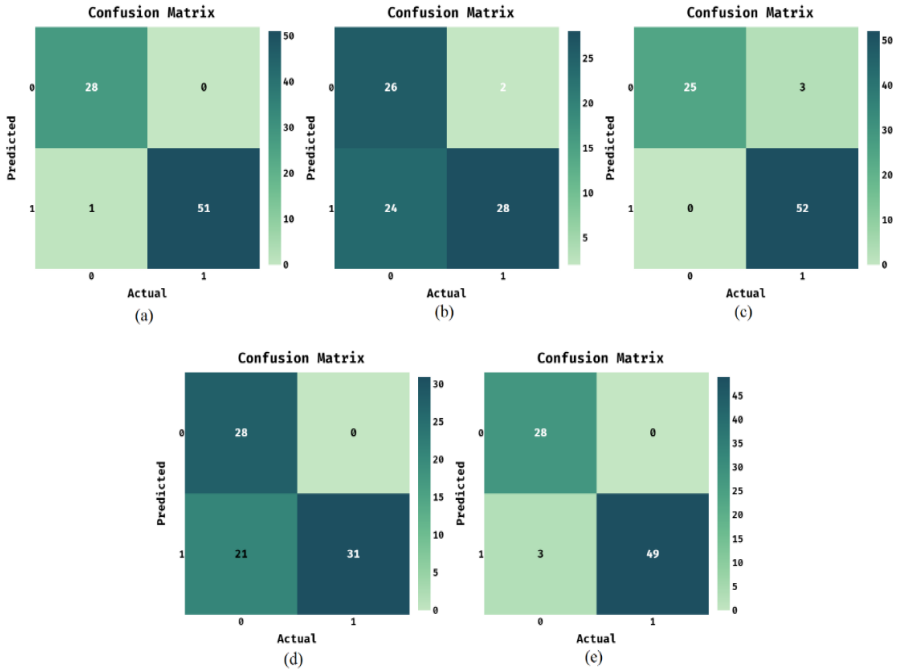

Fig. 5: Confusion Matrix a) RF b) KNN c) DT d) SVM e) LR

Detailed comparative classification results of the ML models take place in Table 2 and Fig. 6. From the resultant values, it is demonstrated that the KNN model has appeared as the worst performer among the ML models which has obtained the lowest precision of 0.79, recall of 0.68, and F-score of 0.68. At the same time, the SVM model has outperformed the KNN model which has attained a slightly increased precision of 0.85, recall of 0.74, and F-score of 0.74. In line with, the DT model has exhibited moderate diagnostic outcome with the precision of 0.96, recall of 0.96, and F-score of 0.96. Meanwhile, the LR model has demonstrated competitive outcome with the precision of 0.97, recall of 0.96, and F-score of 0.96. However, the RF methodology has outperformed superior resultant over the other ML models with the precision of 0.99, recall of 0.99, and F-score of 0.99.

Table 2: Performance Evaluation of Applied Dataset on Various Algorithms with respect to Precision, Recall, F-score

Methods	**Performance Measures**		
	Precision	**Recall**	**F-score**
Random Forest	0.99	0.99	0.99
K Nearest Neighbour	0.79	0.68	0.68
SVM	0.85	0.74	0.74
Decision Tree	0.96	0.96	0.96
Logistic Regression	0.97	0.96	0.96

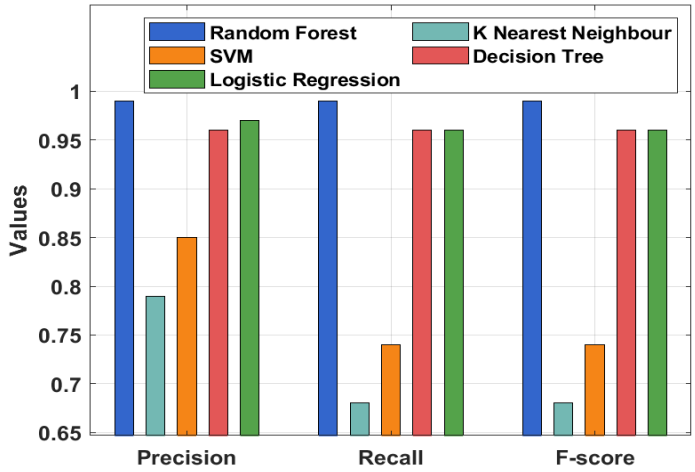

Fig. 6: Result analysis of ML models with different Measures

In order to father validate the classification results of the ML models, another comparative study is made interms of training time, training accuracy, testing accuracy, and error rate in Table 3.Fig. 7 investigates the training accuracy analysis of the ML techniques on the applied CKD dataset. The figure portrayed that the SVM approach has accomplished ineffective results with the minimum training accuracy of 80% whereas the KNN model has gained slightly enhanced performance with the training accuracy of 80.62%. Simultaneously, the LR model has tried to accomplish manageable resultant with a training accuracy of 96.25%. Though the DT technique has offered near optimal outcomes with a training accuracy of 96.25%, the RF model has outperformed all the other methods with the maximum training accuracy of 98.43%.

Table 3: Performance Evaluation of Applied Dataset on Various Algorithms in terms of Various Measures

Methods	**Training Time**	**Training Accuracy**	**Testing Accuracy**	**Error Rate**
Random Forest	0.127651	98.43	98.75	0.012
K Nearest Neighbour	0.093790	80.62	67.50	0.324
SVM	0.010970	80.00	73.75	0.262
Decision Tree	0.156220	97.81	96.25	0.037
Logistic Regression	0.031242	96.25	94.68	0.531

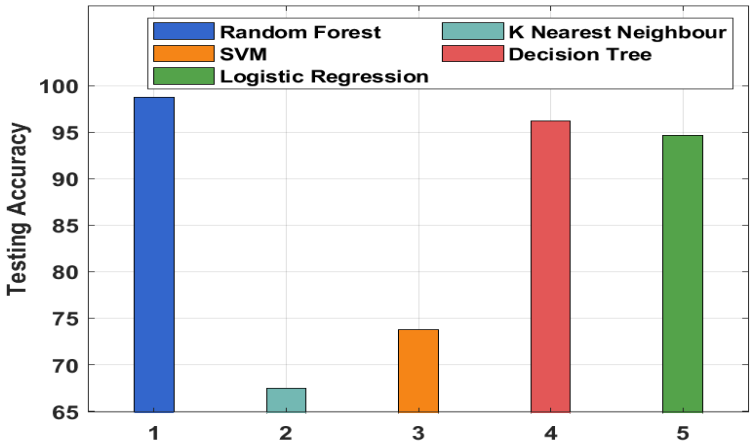

Fig. 7: Training accuracy analysis of ML models

Fig. 8 examines the testing accuracy analysis of the ML techniques on the applied CKD dataset. The figure portrayed that the KNN approach has accomplished ineffective outcomes with the minimum testing accuracy of 67.50% whereas the SVM model has gained slightly enhanced performance with the testing accuracy of 73.75%. Simultaneously, the LR manner has tried to accomplish manageable results with a testing accuracy of 94.68%. Though the DT model has offered near optimal results with a testing accuracy of 96.25%, the RF model has outperformed all the other methods with the maximum testing accuracy of 98.75%.

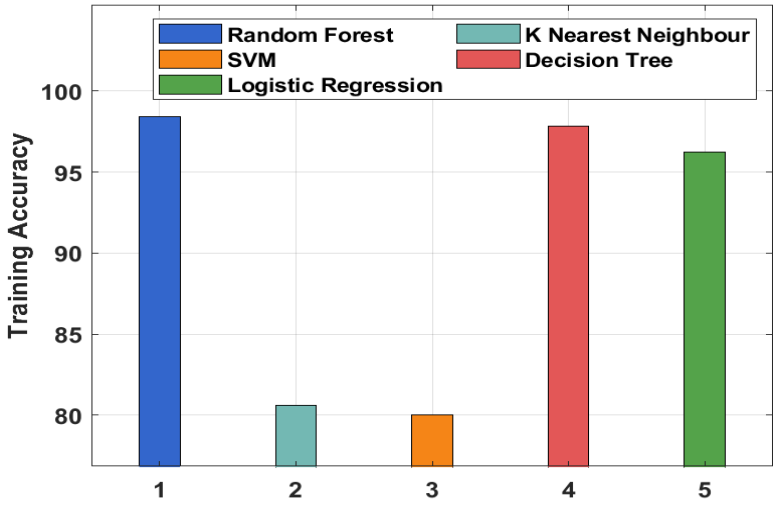

Fig. 8: Testing accuracy analysis of ML models

An error rate analysis of the ML models takes place in Fig. 9 and the value should be low for improved diagnostic performance. From the figure, it can be clear that the LR model has attained least performance with a higher error rate of 0.531. Concurrently, the KNN and SVM models have demonstrated moderately closer error rates of 0.324 and 0.262. Eventually, the DT model has exhibited somewhat reasonable classification performance with a low error rate of 0.037. But the RF model has been found to be effective over the other methods with the lowest error rate of 0.012.

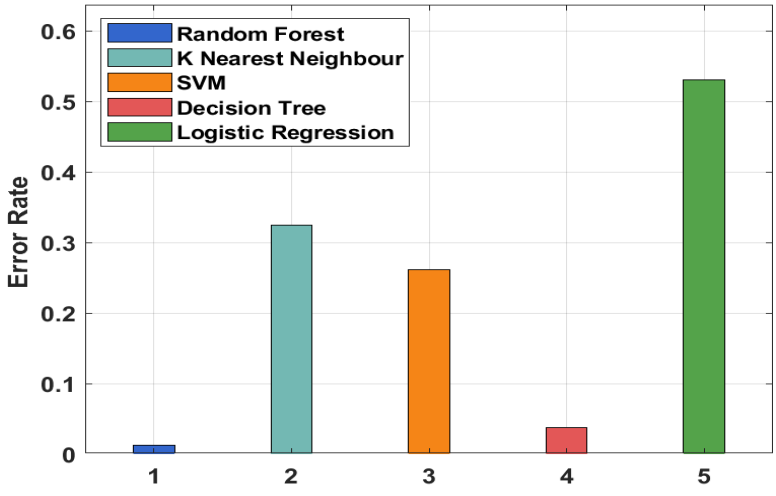

Fig. 9: Error rate analysis of ML models

By observing as to the aforementioned tables and figures, it can be obvious that the RF method has showcased improved classification performance over the KNN, SVM, DT, and LR models. The experimental values stated that the RF model has accomplished a maximum precision of 0.99, recall of 0.99, and F-score of 0.99 with a minimal error rate of 0.012. Therefore, it can be applied as an appropriate tool for CKD diagnosis and classification.

5.0 CONCLUSION

This paper has investigated the CKD diagnostic performance of different ML models. The proposed model initially enables the data pre-processing data to make the input healthcare data compatible for further processing. The data pre-processing involves median based missing value replacement and data transformation. Followed by, five different ML models such as SVM, RF, LR, KNN, and DT are used for data classification. Besides, the performance assessment of these ML models takes place using the benchmark CKD dataset from the UCI repository and the results are examined under different aspects. The experimental outcomes portrayed that the RF classifier is found to be the superior one over the other ML models, which has accomplished a maximum precision of 0.99, recall of 0.99, and F-score of 0.99 with the minimal error rate of 0.012. Therefore, it can be applied as an appropriate tool for CKD diagnosis. In future, the CKD diagnostic performance can be boosted by the use of advanced deep learning architectures.

REFERENCES

[1] S.Oniani, S. et al. "Artificial Intelligence for Internet of Things and Enhanced Medical Systems". *Bio-inspired Neuro computing, Springer, Singapore.* 2021: p. 43-59.

[2] M. Hosseinzadeh, et. al., "A diagnostic prediction model for chronic kidney disease in internet of things platform". *Multimedia Tools and Applications*, 2020. pp.1-18.

[3] M.Elhoseny et. al., "Intelligent diagnostic prediction and classification system for chronic kidney disease". *Scientific reports*, 9(1), 2019, pp.1-14.

[4] J.Uthayakumar, et. al., "Intelligent hybrid model for financial crisis prediction using machine learning techniques", *Information Systems and e-Business Management*, 2018, pp.1-29.

[5] S.K. Lakshmanaprabu., et. al., "Online clinical decision support system using optimal deep neural networks". *Applied Soft Computing*, 81, 2019: p.105487.

[6] P.Koti, et. al., "An efficient healthcare framework for kidney disease using hybrid harmony search algorithm". *Electronic Government, an International Journal*, 16(1-2), 2020:p.56-68.

[7] W.Yu, et. al., "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes",. *BMC medical informatics and decision making*, 10(1), 2010.pp.1-7.

[8] B.Magnin, et. al., "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI", *Neuroradiology*, 51(2), 2009: p.73-83.

[9] C.K. Desai, et. al., "The role of vitamin supplementation in the prevention of cardiovascular disease events", *Clinical cardiology*, 37(9), 2014:p.576-581.

[10] Y.Cao, et. al., "An MLP classifier for prediction of HBV-induced liver cirrhosis using routinely available clinical parameters", *Disease markers*, 35. 2013.

[11] J.Qin, et. al., "A machine learning methodology for diagnosing chronic kidney disease", *IEEE Access*, 8, 2019. 2019: p.20991-21002.

[12] J. Xiao, et. al., *"Comparison and development of machine learning tools in the prediction of chronic kidney disease progression*", Journal of translational medicine, 17(1), 2019.p.1-13.

[13] A. Sobrinho, et. al., *"Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques". IEEE Access*, 8, 2020. Pp.25407-25419.

[14] Jongbo., et.al., "Development of an ensemble approach to chronic kidney disease diagnosis", *Scientific African*, 8, 2020: p.e00456.

[15] Z. Segal, et. al., "Machine learning algorithm for early detection of end-stage renal disease",*BMC nephrology*, 21(1), 2020: p.1-10.

[16] Z.Wang, et. al., "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique". *International Journal of engineering & Technology*, 7, 2017: p.1161-1167.

[17] F. Ma, et. al., "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network", *Future Generation Computer Systems*, 111, 2020: p.17-26.

[18] Qin, J. et. al., "A machine learning methodology for diagnosing chronic kidney disease", *IEEE Access*, 8, 2019: p.20991-21002.

[19] M. Elhoseny, et. al., "Intelligent diagnostic prediction and classification system for chronic kidney disease", *Scientific reports*, 9(1), 2019, p.1-14.

[20] C.F.Chao and M.H. Horng, "The construction of support vector machine classifier using the firefly algorithm", *Computational intelligence and neuroscience*, 2015.

[21] Tian, S. et. al., "Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China", *Remote Sensing*, 8(11), 2016, p.954.

[22] Antonogeorgos. et. al., "Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10-to 12-years-old children: divergence and similarity of the two statistical methods", *International journal of paediatrics*, 2009.

[23] A. Moosavian, et. al., "Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing", *Shock and Vibration*, 20(2), 2013, p.263-272.

[24] Trabelsi, et. al. "Decision tree classifiers for evidential attribute values and class labels", *Fuzzy Sets and Systems, 366*, 2019: p. 1365-1375.