Machines and Algorithms
http://www.knovell.org/mna

Research Article

Classifiers voting based Decision Support System for Prediction of Kidney Related Chronic Diseases

Mubeen Aslam[1, *], Sajid Iqbal[2] and Ahmad Abdullah[2]

[1]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, 54000, Pakistan

[2]Department of Computer Science, Bahauddin Zakariya University, Multan, 60000, Pakistan

[*]Corresponding Author: Mubeen Aslam. Email: mubeen.aslam591@gmail.com

Received: 26 April 2023; Revised: 14 June 2023; Accepted: 20 July 2023; Published: 16 August 2023

AID: 002-02-000022

Abstract: Chronic kidney diseases are increasing exponentially due to hypertension, diabetes, anemia and other related factors. Patients with such diseases usually remain unaware of initial symptoms leading to difficulties in diagnosis of the disease. High performance data mining-based diagnosis and prediction techniques could assist the patient in self-analysis and medical practitioners in developing a precise opinion about patient. This research presents a framework for clinical decision support system of chronic kidney disease (CKD) on the basis of knowledge and facts provided by specialists and experts. To diagnose the disease and decide about progression stage of CKD, different classification algorithms are applied and evaluated on the dataset. The proposed methodology increases the accuracy to 91.75 % and reduces the cost of predicting the stages of CKD using LMT algorithms on the dataset.

Keywords: Chronic kidney disease (CKD); Features mining; Classifier fusion; E-Health;

1. Introduction

Chronic Kidney Disease (CKD) is a durable condition where the kidneys cannot work as expected. According to a study, 1.2 million people died worldwide in 2018 [1], and in Western countries, 5-12% of patients have CKD [2]. Due to the high prevalence of CKD, significant treatment expenses, and inconsistent access to treatment, patients and their families face numerous financial and ethical issues. Creatinine is a crucial measurement for diagnosing CKD; it is a chemical waste product created by muscle metabolism. It's normal range is 1.2mg/dl and 1.46mg/dl for women and men correspondingly however a higher amount of creatinine is produced in the later stages of CKD [3]. Chronic diseases are noncommunicable (NCDs), meaning they do not transfer from one person to another, whereas communicable diseases (CDs) can spread from one person to another and replicate quickly. Chronic disease originates from behavioral, biological, social and environmental factors and can lead to death [8,20]. Such disease can be found throughout the world and among all age groups. The human kidney plays a vital role in the body, and diseases related to it are chronic in nature as well. The major function of a kidney is to filter the blood using millions of nephrons to remove the unwanted chemicals and throw them out of the human body. Non-excretion of these unwanted materials leads to chronic disease in the kidney [9].

Chronic Kidney Disease (CKD) [11] is a resilient disease that has five stages, i.e., CKD stage1 to CKD stage5 and could be diagnosed by several parameters such as high blood pressure, diabetes [10] and anemia. Diabetes increases the sugar level of blood that causes injury to the nerves as well as narrows the vessels [14]. Anemia causes high blood pressure, a shortage of red blood cells (RBCs), and low levels of hemoglobin. Red blood cells provide oxygen to body tissues. The provision of lower oxygen can lead to anemia disease. Anemia decreases iron, red blood cells and changes the shape of RBC. Anemia is a hemoglobin combination with having normal range is less than 12 g/dl and 13 g/dl in women and in men respectively [16]. Another estimate [12] describes that more than hundred peoples per million are affected by kidney diseases alone. According to recent statistics (2019) of the National Kidney Foundation (NKF) [13, 26], USA, the mortality rate of CKD is higher than breast cancer or prostate cancer. It is estimated that only in USA, 37 million people and approximately 90% of those who have CKD don't even know about it. Alarming thing is that around 80 million people are at risk for CKD. Recently, it has been seen that people with kidney disease and transplant recipients are at higher risk for developing serious complications from COVID-19.

To determine the warning signs of CKD, feature-based classification can be performed using machine learning methods. Classification can be done using different attributes of available data with data mining tools to diagnose the disease and extract other relevant information. In order to provide better clinical results to practitioners, expert systems are useful as they can perform automated predictions based on patients' available data. It has been proven in many cases that expert systems can perform better than human specialists due to multidimensional data processing [47, 48]. Various data mining techniques are being used to extract knowledge from existing databases and identify comparative data that could be used in decision-making, assessment, and forecasting [17]. Mostly, descriptive and predictive models are used in data mining. Descriptive models categorize patterns to investigate the properties and relations of data, whereas predictive models predict results from various data sources. Both categories of existing data mining models direct towards various tasks such as prediction, classification, association rule mining, clustering, regression, and time-sequence analysis. They further confirm the actual prediction by using classification and clustering. The state of data can make the problem more complicated; for example, the presence of noise, missing labels, dynamic, and large datasets. Issues with datasets decrease their performance when used in machine learning algorithms.

Similar issues are raised when working with medical datasets to extract unknown patterns and identify the extracted pattern. In the routine diagnosis process, common issues in bioinformatics are not handled properly. Practitioners recommend different test procedures to find out deep information about the disease and formulate their diagnosis. If the set of tests is not formed considering different aspects of the disease, it usually complicates the diagnosis process. Even the use of multiple tests may divert from the correct diagnosis procedure [4], increasing the treatment cost and reducing the performance of prediction methods. These problems can be reduced by using machine learning algorithms that may overcome data deficiencies easily [5]. Classification algorithms are popular in healthcare applications and are used to diagnose and predict the disease in earlier stages [6]. Another aspect that has made classification popular among practitioners is the ability of machine learning methods to deal with complex and large datasets.

Currently, healthcare researchers and industries are applying state-of-the-art statistical methods to assist and guide medical practitioners in treating a wide range of diseases. Statistical methods either use handcrafted features from medical data or automatically extract the features and use them in their decision-making [22]. If the features are handcrafted, the quality of the results depends on the quality of the features used, and the standard of features depends on the knowledge and expertise of the algorithm designer.

The significance of this research is public health impact of CKD, the potential of machine learning in healthcare, and the importance of early detection and intervention in improving patient outcomes in real-time environment.

In this study, we aimed to predict the stages of chronic kidney disease by extracting features from a given dataset. Subsequently, we developed a decision support system to assist both patients and doctors. This system provides information that may not be readily accessible to human experts, without any time

loss. Our goal was to enhance algorithm efficiency and accuracy. To achieve this, we employed a variety of algorithms including logistic model tree, functional tree, J48, Naïve Bayes, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). These algorithms offer valuable insights that can aid physicians in predicting chronic kidney disease at early stages. We utilized a multi-class benchmark dataset and patient history to train the classifiers, fine-tuning various parameters in the process.

The rest of the paper is organized into four sections, Section 2 reports the literature review of different diseases, i.e., diabetes, heart and kidney diseases. Section 3 presents classification methods in detail used in the current era. Section 4, describes the proposed methodology and architecture. Section 5, presents results and discussion, finally, Section 6 concludes research along with future directions.

2. Literature Review

Biological data has experienced significant growth due to multiple factors such as advancements in recording mediums, automated data generation methods, the expansion of medical facilities, and the increase in the human population. Despite this growth, the mortality rate due to various diseases is also on the rise. One major reason for this increased mortality rate is the failure to detect diseases at their initial stages. Some diseases are particularly challenging to diagnose early, including chronic kidney disease (CKD), which progresses slowly alike to kidney failure, cancer, heart disease, asthma, and diabetes. In recent years, numerous classification tasks have been undertaken to predict chronic diseases. Classification algorithms play a crucial role in enhancing the accuracy of disease prediction, with research efforts focused on improving diagnostic accuracy based on clinically collected information. The aim is to detect diseases at early stages to facilitate the development of better treatment options. Popular machine learning algorithms employed in AI-based healthcare systems include Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), and Random Forest (RF).

Diabetes leads to various health complications such as heart disease, kidney failure, blindness, and stroke. Hamedan et al. [21] addresses chronic kidney disease (CKD), emphasizing its subtle symptoms and significant healthcare costs. Three phases were undertaken: identifying variables for the Fuzzy Expert System (FES), developing the FES prototype, and evaluating its robustness with noisy data. Initially, 42 parameters were identified from literature and nephrologist consultation, with seven excluded. Key diagnostic parameters included age, blood pressure, proteinuria, and various biochemical markers. The FES achieved high accuracy 92.12% with demonstrated robustness against noisy data. Additionally, Yadollahpour et al. [15] presents an Expert Medical Decision Support System (MDSS) utilizing an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict the progression of chronic kidney disease (CKD). CKD's covert early stages often delay diagnosis, emphasizing the need for accurate prediction tools to prevent renal damage. The MDSS, based on 10-year clinical records of newly diagnosed CKD patients, predicts Glomerular Filtration Rate (GFR) values, crucial for identifying renal failure. ANFIS, chosen over other models due to its superior accuracy, accurately forecasts GFR variations over 6, 12, and 18-month intervals. The MDSS's performance, evaluated against real patient data, demonstrates high accuracy and efficiency, crucial for improving CKD management and patient outcomes. The user-friendly interface empowers medical professionals with predictive capabilities, enhancing decision-making and patient care. The study underscores the significance of early CKD diagnosis and the potential of ANFIS-based MDSS in improving healthcare outcomes.

Norouzi et al. [28] introduced a medical decision monitoring system for diagnosing kidney failure progression over time. They utilized the Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm to detect kidney failure progression [16], thereby reducing time costs. The dataset, collected from the hospital, consists of 10 attributes used to predict kidney failure. These attributes include age, weight, Glomerular Filtration Rate (GFR), underlying diseases, calcium, creatinine, sex, diastolic blood pressure, phosphorus, and uric acid. The dataset comprises 465 instances, with 277 being male. The reported accuracy of ANFIS is 95%. Charleonnan et al. [9] discussed the application of machine learning techniques in detecting chronic kidney disease (CKD) to aid clinical practices. They employed algorithms such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Decision Tree (DT) to determine the presence of CKD in patients.

The dataset was divided into training and testing sets with a 70:30 ratio. Through five-fold cross-validation, they reported the average accuracy, with SVM achieving the highest accuracy of 98.3%.

To improve the performance of medical systems and reduce mortality rates, Polat et al. [7] proposed an SVM-based method. They utilized online open-source datasets and collected 400 instances with 24 attributes from the UCI machine repository. Feature selection techniques were employed to reduce data dimensionality. The classifier produced binary results predicting the presence or absence of CKD. The authors also utilized SVM with best-first search and achieved 98.5% accuracy using 10-fold cross-validation.

Ahmad et al. [27] addressed various diseases, their symptoms, and major risk factors affecting kidney patients. The study aims to develop a decision support system for doctors to diagnose kidney disease patients. They employed different data mining techniques such as Naïve Bayes, KNN, and Logistic Regression (LR) to predict kidney diseases, demonstrating better performance. Their methodology was based on classification modeling and the development of an expert system. Steps included data collection, preprocessing, and classification, resulting in a reported accuracy of 98.34%. Rodrigues et al. [10] studied the consequences of dialysis treatment, specifically Continuous Ambulatory Peritoneal Dialysis (CAP), for kidney patients. They compiled a dataset containing records of 850 patients over an 8-year period. Naïve Bayes, KNN, Logistic Regression (LR), Multilayer Perceptrons (MLP), and Random Tree (RT) classifier algorithms were employed. K-NN was identified as the best performer among the classifiers, achieving 99.65% accuracy.

Developing specific datasets is a time-consuming and laborious task. While some researchers create their own datasets, many utilize existing datasets provided through open-source licenses. In [49], Subasi et al. conducted binary classification of CKD using an open-source dataset extracted from the UCI online repository. The dataset comprised 400 instances with 24 attributes. Random Forest (RF), ANN, K-NN, and SVM algorithms were employed, with RF achieving the highest performance accuracy at 99.87%.

Another binary classification work related to CKD detection is presented in [45]. Similar to Thiyagaraj et al., the authors obtained their dataset from the UCI online machine learning repository. The dataset includes instances of diabetes, high blood pressure, cardiovascular disease, and family history of kidney failure, categorized into positive and negative features. Preprocessing and clustering techniques were applied for detection and prediction of CKD.

Detection and diagnosis of CKD patients were performed by Mohamed Elhosney et al. in [47]. They collected data from the UCI online repository and applied two classification algorithms: ant-colony-based optimization (D-ACO) and particle swarm optimization (PSO) using 10-fold cross-validation. D-ACO outperformed other methods, achieving 87.5% accuracy. They further analyzed their results based on various aspects such as precision, F-Score, kappa value, and sensitivity, considering the given datasets.

2.1. Comparison in state of the art

To compare our proposed work with existing research, we have defined a set of parameters including dataset size, type of kidney disease, machine learning algorithm used, achieved accuracy, and classification type. Table 1 presents the comparative analysis of existing studies with our proposed work.

Table 1: Comparison of the different dataset with the proposed dataset

Work	Dataset Specification	Kidney Disease Type	ML Algorithms used	Results (accuracy) (Best classifier)	Classification Type
[35]	Instances: 584 Attributes: 6	4-staged kidney disease	Naïve Bayes, Support Vector Machine	SVM with 76.32%.	Multi-class with 5 classes.
[25]	Instances: 584 Attributes: 6	4-staged kidney disease	ANN, Support Vector Machine	ANN with 87.70% accuracy	Multi-class with 5 classes.

[28]	Instances: 465 Attributes: 10	Kidney Failure progression	Adaptive Neuro-Fuzzy Inference system (ANFIS)	ANFIS with 95% accuracy	Binary classification
[9]	Instances: 400 Attributes: 24	Presence or absence of CKD	SVM, Decision Tree, K-NN and Logistic regression	SVM gives higher performance with 98.3%	Binary classification
[7]	Instances: 400 Attributes: 24	Presence or absence of CKD	SVM with wrapped and filtered evaluator in best first search and greedy step wise	SVM filtered best first search gives 98.5% accuracy.	Binary classification
[10]	Instances: 850 Attributes: 8	Kidney dialysis	LR, MLP, Random tree, Naïve Bayes, K-NN	K-NN attains 99.65% accuracy.	Binary Classification
[46]	Instances: 400 Attributes: 25	Presence or absence of CKD	D-ACO and PSO	D-ACO with 87.5% accuracy	Binary Classification
Proposed dataset	Instances: 800 Attributes: 25	5-stages CKD	ANN, SVM, Naïve Bayes, J48, LMT, FT	LMT has a greater accuracy of 91.375% than other classifiers.	Multi-class with 6 classes

*Stage 1 CKD: eGFR 90 or Greater, Stage 2 CKD: eGFR Between 60 and 89, Stage 3 CKD: eGFR Between 30 and 59, Stage 4 CKD: eGFR Between 15 and 29, Stage 5 CKD: eGFR Less than 15

Table 1 illustrates that the proposed methodology outperforms other works in several aspects. For instance, Vijayarani et al. [35] diagnosed different stages of kidney diseases using 584 instances with 6 parameters and employed Naïve Bayes and SVM classifier algorithms. They achieved a greater accuracy of 76.32% in multi-class classification among 5 classes using SVM. Similarly, Vijayarani et al. [35] classified the multi-class and diagnosed different stages of kidney disease with ANN, achieving a performance of 87.70%. Polat et al. [7] conducted binary classification to predict chronic kidney disease using SVM with wrapped and filtered evaluator, achieving a performance accuracy of 98.5% with SVM filtered best first search evaluator. Mohamed et al. suggested a binary classification method to determine if a patient has CKD or not, with D-ACO providing an accuracy of 87.5% compared to other algorithms. Lakshimi et al. [23] collected data from different dialysis sites to detect kidney dialysis, analyzing it using three data mining classifier algorithms: ANN, decision tree, and logistic regression model. ANN emerged as the highest performer with 93.853% accuracy, using binary classes 'survive' and 'die' for classification, predicting patient survivability by corresponding class values. Norouzi et al. [28] addressed the Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier algorithms, demonstrating its utility in diagnosing kidney failure progression over time. They exploited the binary class of kidney failure progression, achieving a prediction accuracy of 95%. Charleonnan et al. recommended an approach to predict the chronic kidney disease state of a patient with binary classes 'CKD' and 'not CKD', utilizing K-NN, SVM, decision tree, and logistic regression. However, our proposed methodology employs multi-classification and utilizes a larger dataset with more features, resulting in better performance than other works performing multi-classification. SVM provided a higher accuracy of 91.375% among 25 attributes, correctly diagnosing the chronic kidney disease stages.

3. Classification

The research focuses on chronic kidney disease (CKD), highlighting its global prevalence and significant impact on public health. To address this, the study emphasizes the importance of early detection

and intervention. Utilizing machine learning algorithms like Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Trees (J48), Naïve Bayes (NB), Logistic Model Trees (LMT), and Functional Trees (FT), the research aims to develop accurate diagnostic tools for CKD. These algorithms enable efficient pattern recognition and classification, offering promising ways for improving patient care through timely interventions.

3.1. Artificial Neural Network (ANN)

Artificial Neural Networks and deep learning algorithms have become the state of the art in Artificial Intelligence applications for pattern recognition. ANNs are collections of a large number of neurons connected with each other in a defined way. There are numerous successful applications of ANNs in medical data to solve various problems like image analysis, drug development, interpretation, and prediction. Successful implementation of ANN-based algorithms provides more confidence to clinical practitioners as well as researchers about the achieved results. ANNs operate either in a cascaded or hierarchical way where results produced by one set of neurons are propagated to the next set. Usually, an ANN has multiple layers divided among input layer, hidden layers, and output layer. Figure 1 illustrates the general configuration of an ANN.

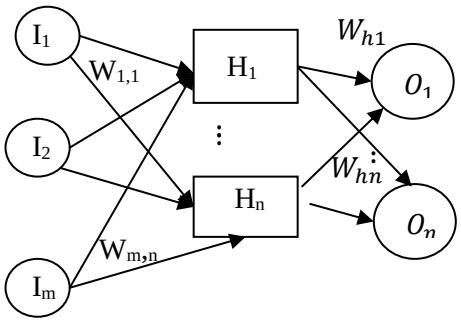

Figure 1: General ANN architecture

The input layer takes data from the environment in the form of numbers which can represent any data type like image, text, numeric or even speech. The input neurons are next connected to hidden neurons and next layer neurons (hidden layer) computes the hyper parameters for each connection as shown in the figure. The most important hyper parameters are the weights which are assigned to each link between the two layers. In figure 1, weights are denoted by W. Each next layer neuron computes the sum of products using following equation-1.

$$H_j = \sum_{i=1}^{n} I_i W_{ij} + b_j \tag{1}$$

where,

I_i = The input coming through i^{th} input neuron

J = Neuron index in hidden layer

Similarly, each output can be computed using equation-2.

$$O_j = \sum_{i=1}^{n} H_i W_{ij} + b_j \tag{2}$$

Any neuron other than input layer consists of two parts: summation and activation function that either filters or smooths the coming input. An activation function could be linear or non-linear and normally boosts the performance of classifier. ANN mostly use the supervised learning where output produced by the final layer neurons are compared with the actual target values and the difference of the predicted and actual target values, known as loss, is calculated. Based on calculated loss, the hyper-parameters (weights and biases) are fine-tuned such that they minimize the loss. The said process takes large number of iterations.

3.2. Support Vector Machine (SVM)

SVM is widely used in machine learning (ML) and pattern recognition applications due to its high performance compared to other ML methods. It is a supervised learning technique utilized for regression and classification tasks. Linear SVM supports the use of a hyperplane to separate the two classes of data. Its binary class classification capability can be readily extended for multiclass classification. SVM utilizes support vectors, which endeavor to maximize the margin from the hyperplane, as illustrated in Figure 2.

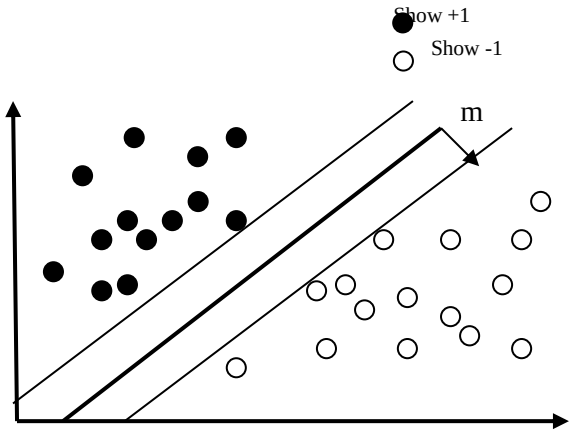

Figure 2: Linear SVM separation hyperplane

Nearest points to the margin line are called the support vector points and help in determine the optimal boundary for given classes. SVM can perform both linear and non-linear boundary detection. The negative plane represents less than 1 value in the SVM technique. It is given by equation-3.

$$z_i = m.x_i + h \leq -1 \tag{3}$$

The positive plane represents the greater than 1value in SVM technique. It is given by equation.

$$z_i = m.x_i + h \geq 1 \tag{4}$$

The classifier boundary of hyperplane is given by equation 5.

$$z_i = m.x_i + h \tag{5}$$

Nonlinearly separable hyperplane is more powerful than its linear counterpart.

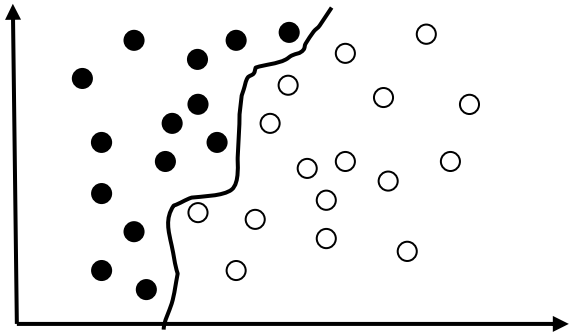

Figure 3: Non-linear SVM hyperplane

By combining both equations (3) and (4)

$$z_i(m.x_i + h) \geq 0 \ \forall_i \tag{6}$$

Quadratic Programming problem that arises during the training of SVM is solved using Sequential Minimum Optimization (SMO) algorithm [30] and optimizes the performance. SMO breaks the problem into small chunks (sub-problem), which are solved systematically one by one in a series.

3.3. J48 Decision Tree

J48 is a decision tree classifier, which is essentially a Java-based implementation of the C4.5 method. It utilizes information theory to evaluate the individual features of a given dataset. Using information gain, it determines the best split in the decision tree. The objective of J48 (Weka implementation of C4.5) is to achieve decision accuracy with flexibility. Decision tree pruning is the process that eliminates some tree nodes/branches without affecting the accuracy of the model. This removal of branches reduces the size of the tree and enhances computational efficiency. Another benefit of tree pruning is to prevent overfitting during training. J48 employs two methods of tree pruning: subtree replacement and subtree raising. Some studies have demonstrated that the pruning process can also improve the efficiency of the decision tree algorithm.

3.4. Naïve Bayesian (NB)

The Naïve Bayes algorithm is a supervised and probabilistic classifier. NB specifies conditional independence among attributes. Its simplicity lies in the simple multiplication of probabilities, reducing complexity. Due to its straightforward nature, this classification method is rapidly adopted. It achieves accurate parameter estimation by calculating the frequencies of attributes and combinations of values in a given training dataset. The algorithm has been found to be equally efficient for medical data. Despite the assumption of attribute independence, the NB classifier generates better accuracy performance. If the probability of a given data instance lying in class i is denoted by v_i, then for n classes, probabilities are given by $V = \{v\_1, v\_2, \ldots, v\_n\}$. The probability for a given instance to lie in specified classes is given by:

$$P(v_j) = \sum_{i=1}^{n} P(v_j|C_i)P(C_i) \tag{7}$$

3.5. Logistic Model Tree (LMT)

LMT is an amalgamation of the decision tree and logistic regression model. The decision tree is the most conventional method for classification, where each instance is classified based on its parameter values. It subdivides instances to determine the particular class, depending on the different values of the given parameters. Using information gain theory, the most useful attribute is selected first to define the tree. The node with the highest rank is chosen as the root of the tree. The decision tree is a subdivision of tree nodes by splitting every instance until it finds the class label. It holds a non-linear model to classify data with easy

interpretation and is preferable for a small amount of training data. Entropy D is used to measure the data impurity. For a dataset of instances $I = \{I_1, I_2, I_3, \dots, I_n\}$, entropy is given as:

$$Entropy(D) = \sum_{i=1}^{n} -I_i log I_i \tag{9}$$

And information gain for a particular feature is given by:

$$InformationGain(S) = entropy(D) - \sum_{i=1}^{m} \frac{|D_i|}{|D|} entropy D_i \tag{10}$$

The logistic regression captures the linear patterns to classify the data which portrays the data with low variance and high bias. Logistic model is preferred when there are small number of instances with noise.

Logistic regression model separates each class from parent class 'K' by using 'K-1' log odds:

$$\beta_k^T z = \log\left(\frac{P(G = k|Z = z)}{P(G = j|Z = z)}\right) \tag{11}$$

For $k = 1, \dots\dots, K-1$, β_k and P_i is the estimation of the attribute values and probability of linear function z respectively. The performance of both approaches depends on the number of instances and number of attributes of the dataset [40, 41] with no global ranking. However, LMT gives better results because it is the mixture of a logistic regression with decision tree and logistic regression is applied to the leaves of the tree using Logit Boost [42,43]. In LMT, the decision tree makes a tree that divides the set of instances into 3 regions and applies the logistic regression in every region. When both methods apply on a dataset, its performance is increased as compared to simple decision tree classifier and simple logistic regression models. LMT increases the computational complexity due to the regression function that is applied to the tree leaves.

3.6. Functional Tree

The functional tree builds the decision tree and applies logistic regression at nodes and/or leaves [44]. The decision tree is built in two phases in the functional tree algorithm. Traversing the decision tree in a depth-first tree for pruning the decision tree. In the first phase, a decision tree is built and pruning of the decision tree is done in the second phase. For pruning the tree two measures are estimated at non-leaf node.

4. Proposed Methodology

The research proposes a multiphase approach to developing a decision support system (DSS) for chronic kidney disease (CKD) diagnosis, breaking down the process into specific phases to ensure thoroughness and systematic development. It integrates data mining and machine learning techniques within the DSS framework to enhance diagnostic accuracy and performance, representing a novel approach to CKD diagnosis. Utilizing a specialized dataset sourced from a healthcare unit ensures the relevance and applicability of the data to real-world scenarios. Advanced feature ranking and selection methods prioritize high-value attributes crucial for assessing kidney function, potentially improving the DSS's accuracy. The research highlights the significance of gender-specific differences in physiological parameters for CKD diagnosis, acknowledging the importance of personalized medicine. Through experimental results analysis, the research provides insights into the performance of different classification algorithms, informing future research and clinical applications. The structure of proposed DSS is shown in figure-4.

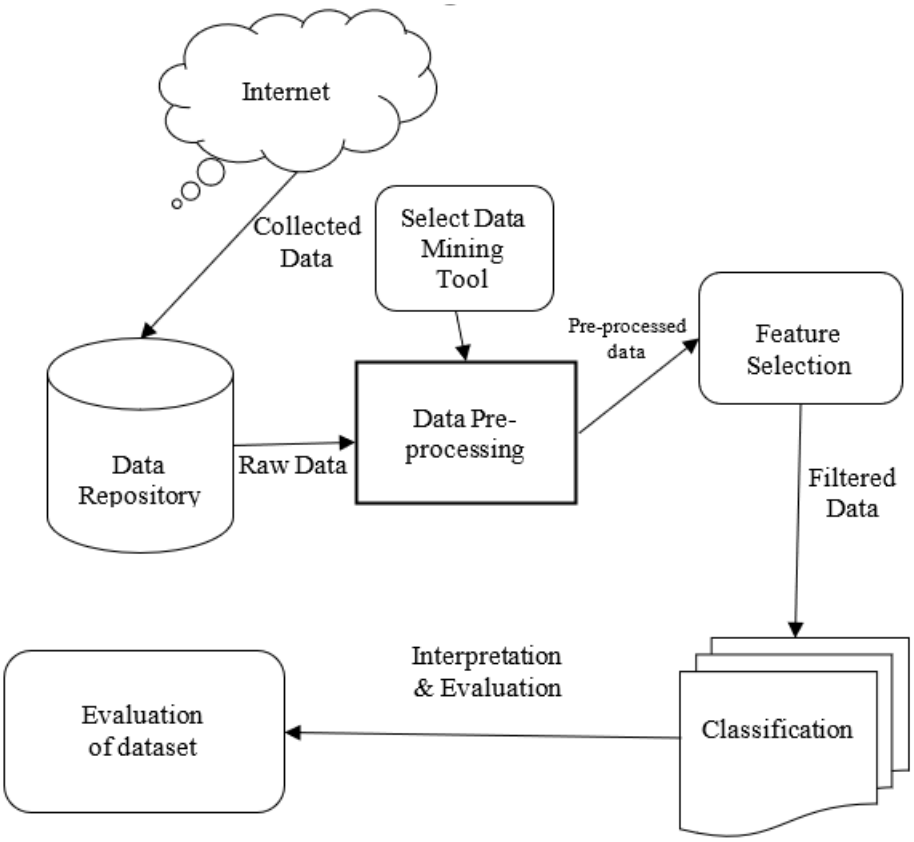

Figure 4: Flow Diagram for performance evaluation

4.1. Data collection

The data is sourced from the online machine learning repository of the University of California at Irvine, accessible at https://archive.ics.uci.edu/ml/index.php. Chronic disease dataset was collected from a healthcare unit over a period of 2 months. The dataset is provided by Dr. P. Soundarapandian, M.D., D.M., Senior Consultant Nephrologist, Apollo Hospitals, Managiri, Madurai Main Road, Karaikudi, Tamilnadu, India. It is a multivariate dataset specifically designed for classification tasks, comprising 800 instances with 25 attributes, including 11 numeric and 14 nominal attributes with multi-labels. The dataset is labeled with 5 CKD stages and one normal state. The extracted features include age, weight, gender, blood pressure, specific gravity, presence of diabetes, albumin, red blood cells, pus cells, pus cell clumps, amount of single-cell bacteria, random blood glucose, urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, hypertension, diabetes mellitus, presence of coronary artery disease, appetite, pedal edema, anemia, and the class of the instance. The five stages are CKD Stage 1: eGFR 90 or Greater, CKD Stage 2: eGFR Between 60 and 89, CKD Stage 3: eGFR Between 30 and 59, CKD Stage 4: eGFR Between 15 and 29, CKD Stage 5: eGFR Less than 15. The training dataset contains missing values and noise for some instances. Another test dataset was collected from local healthcare units. For classification, 10-fold cross-validation with a 90:10 ratio is performed on the formulated dataset. To measure the performance of classification algorithms, a separate new dataset (named TEST) is prepared by collecting relevant data from local hospitals and clinics. This test dataset has 66 instances and the same number of features.

4.2. Data mining Tool Selection

To apply the data mining algorithms and relevant data processing procedures, a data mining tool Weka 3.6 is selected. The selection is made considering its features. The tool has a number of built-in data mining

algorithms with data preprocessing features. The tool also provides the data visualization and evaluation modules.

4.3. Preprocessing

In this phase, noise and missing values are removed from dataset. Missing values are replaced with mean values. Different attributes of dataset contain values in different ranges, all such attributes are standardized in the range 0-1.

4.4Feature Ranking

The next step in diagnosis is feature ranking, where high-value features are utilized first in the mining process, while low-value features are exposed to the classifier at a later stage. For this purpose, information gain and entropy theory are employed. The highest-ranked attributes include Serum Creatinine, blood urea, hemoglobin, specific gravity, hypertension, red blood cell count, and diabetes mellitus, which manage the risk factors of kidney function. Serum Creatinine and blood urea gain high ranks because they indicate the working condition of the kidney and serve as early signs of kidney malfunction. Hemoglobin represents the rate of red blood cells in human beings, while specific gravity measures the kidney's ability to concentrate urine with plasma. Hypertension is a dangerous sign of chronic kidney disease, as it increases the risk factor of kidney failure. A smaller number of red blood cell counts in CKD patients also heightens the risk of kidney malfunction.

4.5. Classification

Classification is two step processing: feature mapping of labels and application of classification algorithm also known as statistical model. The classifier is trained using train dataset and its performance is evaluated using test/validation dataset. A number of classification methods are used in this work and their performance is reported in terms of precision, recall and f-measure. In our experiment, we first performed the binary classification and then multi-classification. The results for binary-classification are listed in table 3

Table 3: Binary class dataset results

Algorithms	Accuracy (%)	Test time (second)
LMT	98%	0.99s
FT	97.5%	0.09s
J48	99%	0.01s
ANN	99.75%	5.05s
SVM	97.75%	0.02s
Naïve Bayes	95%	0.01s

Table 3 characterizes the binary dataset classification results in terms of accuracy and test time in seconds. In binary classification, the ANN algorithm achieves 99.75% accuracy, albeit consuming more time compared to other algorithms. Naïve Bayes takes 0.01s to provide a 95% accuracy rate. Although the ANN algorithm utilizes slightly more time than other classification algorithms, it offers higher accuracy. ANN outperforms decision tree-based algorithms (LMT, FT, J48) in binary classification because the decision tree selects specific attributes from the dataset and performs classification on those attributes. However, the ANN algorithm uses all the attributes selected during preprocessing. Naïve Bayes, being a simple statistical model (non-parametric), shows lower performance as it cannot exploit the non-linearity in data at a deeper level. The decision tree (LMT, FT, J48) classifier achieves superior accuracy with insignificant differences.

In our second experiment, we conducted multi-classification for the selected dataset. Results obtained for multi-classification are listed in Table 4.

Table 4: Multi-class dataset-2 results

Algorithms	Accuracy (%)	Test Time (second)
LMT	84.5%	4.91s
FT	84.5%	0.2s
J48	84.75%	0.01s
ANN	66.75%	6.22s
SVM	67%	0.2s
Naïve Bayes	77.75%	0.1s

The J48 algorithm achieves 84.75% accuracy in 0.01 seconds. LMT and FT also yield better results with slightly lower accuracy, at 84.5% in 4.91 seconds and 0.2 seconds, respectively. Dataset-2 focuses solely on identifying CKD stages without utilizing the gender attribute. However, we observed that the gender attribute holds significant importance for CKD stages. This is because certain selected features, such as serum creatinine and hemoglobin lab reports, exhibit different normal values based on gender.

Dataset-3 includes 27 attributes, including the gender attribute. To balance the classes, an equal number of instances are added for both genders.

Table 5: Multi-class dataset-3 results

Algorithms	Accuracy (%)	Test Time (second)
LMT	91.75%	10.65s
FT	87.5%	0.49s
J48	85.875%	0.03s
ANN	76.625%	14s
SVM	75.5%	0.26s
Naïve Bayes	81.5%	0.01s

In Table 5, LMT demonstrates greater performance than SVM, primarily attributed to the inclusion of the high-ranking feature 'gender'. LMT achieves a higher accuracy than ANN as well. Specifically, LMT yields the best result with a 91.75% accuracy rate in 10.65 seconds, while SVM consumes less time (0.26 seconds) but achieves a lower accuracy of 75.5%. Based on the outcomes of the three experiments, we draw the following conclusions: In the first experiment, the results are notably high due to binary classification, which is a comparatively easier task. However, in the second experiment, the results experience a drastic drop due to the dataset's multi-classification nature. Nevertheless, in the third experiment, there is a noticeable improvement in the results, attributed to the inclusion of high-value features.

By following the outlined methodology, doctors can utilize advanced data mining and machine learning techniques to enhance their diagnostic capabilities. The DSS facilitates the integration of patient data from various sources, enabling clinicians to identify key biomarkers and clinical indicators associated with CKD progression. Through sophisticated feature ranking and selection methods, doctors can prioritize relevant attributes essential for accurate diagnosis, such as serum creatinine levels, blood pressure, and demographic factors. This systematic approach empowers healthcare professionals to make informed decisions based on evidence-driven insights extracted from patient data. Additionally, the DSS provides real-time monitoring

and predictive analytics capabilities, enabling doctors to anticipate CKD progression and implement timely interventions to mitigate adverse outcomes.

5. Evaluation and Result

Cross validation is the method to produce results with high confidence. As stated above, we used 10-fold cross validation for our experiments and we report the average score. To be surer about our results, we used another dataset for test purpose which was collected from local hospitals with the help of relevant medical practitioners.

5.1. Experiment-1

Here, we use recall, accuracy and f-measure and precision to report results of our experiments.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

$$f - measure = \frac{2TP}{2TP + FP + FN} \tag{14}$$

$$Preceision = \frac{TP}{TP + FP} \tag{15}$$

ANN performed best for experiment-1 and its confusion matrix is given below in table 6.

Table 6: Confusion Matrix for CKD Class

Predicted Value	**Actual**		**Performance measures**			
	Positive	**Negative**	**Recall**	**Accuracy**	**f-measure**	**Precision**
Positive	498	0	0.966	99.75%	0.998	1
Negative	2	300				

The accuracy is higher for experiment-1 and it is due to the fact that we are doing binary classification about CKD which is relatively easy problem.

5.2. Experiment-2

In experiment-2, we performed the multi-classification. In this experiment the dataset used contains all features except gender. Among all classifiers, J48 performed best and following table 8 shows its confusion matrix.

Table 7: Confusion Matrix for multi-classification

Predicted Value per CKD stage	Actual						Performance measures			
	1	**2**	**3**	**4**	**5**	**0**	**Recall**	**Accuracy**	**f-measure**	**Precision**
1	46	12	2	4	0	8	0.76	84.5	0.86	0.99
2	10	48	14	0	0	2				
3	4	16	134	6	0	2				
4	2	2	16	68	8	0				
5	0	0	0	12	84	0				
0	0	4	0	0	0	296				

The J48 algorithm achieves the best result for dataset-2 with an 84.5% performance accuracy. The description of each stage is specified as follows:

- For CKD stage 1 and CKD stage 2 patients, the age and weight differ, while the serum creatinine value remains the same.
- CKD stage 3 is determined by the serum creatinine and RBCC (Red Blood Cell Count) attribute values, with the RBCC value being less than 4.3 million/cm.
- For CKD stage 4, the serum creatinine value is greater than or equal to 1.6 mg/dl.
- The attributes values for CKD stage 5 are very close to those of CKD stage

5.3. Experiment-3

In this experiment, we added the gender attribute in the dataset and performed the multi-classification. Following table 8 shows the confusion matrix and results obtained for high performing method i.e., LMT.

Table 8: Confusion Matrix for multiclassification on dataset-3

Predicted Value per CKD stage	Actual						Performance measures			
	1	**2**	**3**	**4**	**5**	**0**	**Recall**	**Accuracy**	**f-measure**	**Precision**
1	53	7	0	0	0	0	0.868	91.75	0.93	1
2	9	60	5	1	0	1				
3	0	10	143	5	0	0				
4	0	3	12	82	3	0				
5	0	0	0	10	96	0				
0	0	0	0	0	0	300				

Table 8 shows that dataset-3 gives better performance than dataset-2 for each stage of CKD. Serum creatinine, gender, age, and weight are required to find the stages of CKD but all the other attributes that are used in the dataset are significant to predict and control the CKD stages. Serum creatinine, RBCs state, number of RBCC, anemia sign and hemoglobin attributes seen in CKD stage 5 patients in critical condition. Every stage of CKD is very close to the other.

5.4. Classification evaluation using proprietary test dataset

A separate test dataset (TEST) is used to validate the results of our trained models by using LMT algorithm. The results obtained using this dataset are given in table 9.

Table 9: Confusion Matrix for multi-classification on TEST

Predicted Value per CKD stage	**Actual**						**Performance measures**			
	1	**2**	**3**	**4**	**5**	**0**	**Recall**	**Accuracy**	**f-measure**	**Precision**
1	10	0	0	0	0	1	0.864	86.36	0.913	1
2	0	5	3	0	0	0				
3	0	1	9	0	0	2				
4	0	0	0	11	2	0				
5	0	0	0	0	12	0				
0	0	0	0	0	0	10				

By utilizing advanced data mining and machine learning techniques as outlined in the developed DSS framework, clinicians significantly enhance their diagnostic capabilities for chronic kidney disease (CKD). The experiments demonstrate the DSS's efficacy in accurately diagnosing CKD and predicting the disease stage, leveraging patient data from diverse sources. Through sophisticated feature ranking and selection methods, the DSS enables clinicians to identify crucial biomarkers and clinical indicators associated with CKD progression. Moreover, the DSS provides real-time monitoring and predictive analytics, empowering doctors to anticipate disease progression and implement timely interventions.

6. Conclusion and Future Work

In conclusion, this research proposes a comprehensive framework for the clinical decision support system (DSS) of chronic kidney disease (CKD) diagnosis and progression prediction. By utilizing high-performance data mining techniques and classification algorithms, the study aimed to improve the accuracy and efficiency of CKD diagnosis, enabling both patients and medical practitioners to make informed decisions. The findings indicate that the proposed methodology significantly enhances the diagnostic capabilities for CKD, particularly in early detection and disease progression prediction. By prioritizing high-value features and advanced machine learning techniques, the developed DSS demonstrates its potential to assist healthcare professionals in making timely and accurate diagnoses, thereby improving patient outcomes and reducing the burden of CKD-related complications. The continued refinement and validation of the proposed DSS framework, along with the exploration of emerging technologies and methodologies, hold promise for advancing the diagnosis and management of CKD and improving patient care in the future.

References

[1] Luyckx, Valerie A., Marcello Tonelli, and John W. Stanifer. "The global burden of kidney disease and the sustainable development goals." *Bulletin of the World Health Organization* 96, no. 6 (2018): 414.

[2] Versino, Elisabetta, and Giorgina Barbara Piccoli. "Chronic kidney disease: the complex history of the organization of long-term care and bioethics. Why now, more than ever, action is needed." *International Journal of Environmental Research and Public Health* 16, no. 5 (2019): 785.

[3] Pandya, Divya, Anil Kumar Nagrajappa, and K. S. Ravi. "Assessment and correlation of urea and creatinine levels in saliva and serum of patients with chronic kidney disease, diabetes and hypertension–a research study." *Journal of clinical and diagnostic research: JCDR* 10.10 (2016): ZC58.

[4] Muthulakshmi, I. "An Extensive Survey on Evolutionary Algorithm Based Kidney Disease Prediction." In *2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*, pp. 1-5. IEEE, 2019.

[5] Kumar, Manish. "Prediction of chronic kidney disease using random forest machine learning algorithm." *International Journal of Computer Science and Mobile Computing* 5, no. 2 (2016): 24-33.

[6] www.nhlbi.nih.gov/health/health-topics/topics/hbp.

[7] Polat, Huseyin, Homay Danaei Mehr, and Aydin Cetin. "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods." *Journal of medical systems* 41 (2017): 1-11.

[8] Bala, Suman, and Krishan Kumar. "A literature review on kidney disease prediction using data mining classification technique." *International Journal of Computer Science and Mobile Computing* 3, no. 7 (2014): 960-967.

[9] Charleonnan, Anusorn, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, and Nitat Ninchawee. "Predictive analytics for chronic kidney disease using machine learning techniques." In *2016 management and innovation technology international conference (MITicon)*, pp. MIT-80. IEEE, 2016.

[10] Rodrigues, Mariana, Hugo Peixoto, Marisa Esteves, and José Machado. "Understanding stroke in dialysis and chronic kidney disease." *Procedia computer science* 113 (2017): 591-596.

[11] Neves, José, M. Rosário Martins, João Vilhena, João Neves, Sabino Gomes, António Abelha, José Machado, and Henrique Vicente. "A soft computing approach to kidney diseases evaluation." *Journal of medical systems* 39 (2015): 1-9.

[12] Hill, Nathan R., Samuel T. Fatoba, Jason L. Oke, Jennifer A. Hirst, Christopher A. O'Callaghan, Daniel S. Lasserson, and FD Richard Hobbs. "Global prevalence of chronic kidney disease–a systematic review and meta-analysis." *PloS one* 11, no. 7 (2016): e0158765.

[13] https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics#:~:text=1%20in%203%20American%20adults,lived%20with%20a%20kidney%20transplant.

[14] American Diabetes Association. "9. Cardiovascular disease and risk management: standards of medical care in diabetes—2018." *Diabetes care* 41, no. Supplement_1 (2018): S86-S104.

[15] Karthikeyan, T., and P. Thangaraju. "Analysis of classification algorithms applied to hepatitis patients." *International Journal of Computer Applications* 62, no. 15 (2013): 2530.

[16] Karthikeyan, T., and P. Thangaraju. "Best first and greedy search based CFS-Naïve Bayes classification algorithms for hepatitis diagnosis." *Biosciences and Biotechnology Research Asia* 12, no. 1 (2015): 983-990.

[17] Mittal, Ansh, Deepika Kumar, Mamta Mittal, Tanzila Saba, Ibrahim Abunadi, Amjad Rehman, and Sudipta Roy. "Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images." *Sensors* 20, no. 4 (2020): 1068.

[18] Iraji, Mohammad Saber. "Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing." *Journal of Applied Biomedicine* 15, no. 2 (2017): 151-159.

[19] Kinaan, Mustafa, Hanford Yau, Suzanne Quinn Martinez, and Pran Kar. "Concepts in Diabetic Nephropathy: From Pathophysiologyto Treatment." *Journal of Renal and Hepatic Disorders* 1, no. 2 (2017): 10-24.

[20] Webster, Angela C., Evi V. Nagler, Rachael L. Morton, and Philip Masson. "Chronic kidney disease." *The lancet* 389, no. 10075 (2017): 1238-1252.

[21] Das, Himansu, Bighnaraj Naik, and H. S. Behera. "Classification of diabetes mellitus disease (DMD): a data mining (DM) approach." In *Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017*, pp. 539-549. Springer Singapore, 2018.

[22] Mohapatra, Subasish, Prashanta Kumar Patra, Subhadarshini Mohanty, and Bhagyashree Pati. "Smart health care system using data mining." In *2018 International Conference on Information Technology (ICIT)*, pp. 44-49. IEEE, 2018.

[23] Fan, Li, Andrew S. Levey, Vilmundur Gudnason, Gudny Eiriksdottir, Margret B. Andresdottir, Hrefna Gudmundsdottir, Olafur S. Indridason, Runolfur Palsson, Gary Mitchell, and Lesley A. Inker. "Comparing GFR estimating equations using cystatin C and creatinine in elderly individuals." *Journal of the American Society of Nephrology* 26, no. 8 (2015): 1982-1989.

[24] Borisagar, Nilesh, Dipa Barad, and Priyanka Raval. "Chronic kidney disease prediction using back propagation neural network algorithm." In *Proceedings of International Conference on Communication and Networks: ComNet 2016*, pp. 295-303. Springer Singapore, 2017.

[25] Vijayarani, S., S. Dhayanand, and M. Phil. "Kidney disease prediction using SVM and ANN algorithms." *International Journal of Computing and Business Research (IJCBR)* 6, no. 2 (2015): 1-12.

[26] En Espanol, Chronic Kidney Diseases *http://www.kidney.org/kidneydisease/aboutckd.cfm*

[27] Ahmad, Mubarik, et al. "Diagnostic decision support system of chronic kidney disease using support vector machine." *2017 Second International Conference on Informatics and Computing (ICIC)*. IEEE, 2017.

[28] Norouzi, Jamshid, Ali Yadollahpour, Seyed Ahmad Mirbagheri, Mitra Mahdavi Mazdeh, and Seyed Ahmad Hosseini. "Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system." *Computational and mathematical methods in medicine* 2016, no. 1 (2016): 6080814.

[29] De Guia, Justin D., Ronnie S. Concepcion, Argel A. Bandala, and Elmer P. Dadios. "Performance comparison of classification algorithms for diagnosing chronic kidney disease." In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pp. 1-7. IEEE, 2019.

[30] Virkar, Hemant, Karen Stark, and Jacob Borgman. "Machine learning method that modifies a core of a machine to adjust for a weight and selects a trained machine comprising a sequential minimal optimization (SMO) algorithm." U.S. Patent 9,082,083, issued July 14, 2015.

[31] Rehman, Amjad, Naveed Abbas, Tanzila Saba, Syed Ijaz ur Rahman, Zahid Mehmood, and Hoshang Kolivand. "Classification of acute lymphoblastic leukemia using deep learning." *Microscopy Research and Technique* 81, no. 11 (2018): 1310-1317.

[32] Ramzan, Farheen, Muhammad Usman Ghani Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks." *Journal of medical systems* 44 (2020): 1-16.

[33] Rehman, Amjad, Muhammad A. Khan, Zahid Mehmood, Tanzila Saba, Muhammad Sardaraz, and Muhammad Rashid. "Microscopic melanoma detection and classification: A framework of pixel-based fusion and multilevel features reduction." *Microscopy research and technique* 83, no. 4 (2020): 410-423.

[34] Sharif, Uzma, Zahid Mehmood, Toqeer Mahmood, Muhammad Arshad Javid, Amjad Rehman, and Tanzila Saba. "Scene analysis and search using local features and support vector machine for effective content-based image retrieval." *Artificial Intelligence Review* 52 (2019): 901-925.

[35] Vijayarani, S., and S. Dhayanand. "Data mining classification algorithms for kidney disease prediction." *Int J Cybernetics Inform* 4, no. 4 (2015): 13-25.

[36] Ahishakiye, Emmanuel, Danison Taremwa, Elisha Opiyo Omulo, and Ivan Niyonzima. "Crime prediction using decision tree (J48) classification algorithm." *International Journal of Computer and Information Technology* 6, no. 3 (2017): 188-195.

[37] Quinlan, John R. "Learning with continuous classes." In *5th Australian joint conference on artificial intelligence*, vol. 92, pp. 343-348. 1992.

[38] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." *arXiv preprint arXiv:1206.1121* (2012).

[39] Ahmed, Fahim, and Kyoung-Yun Kim. "Data-driven weld nugget width prediction with decision tree algorithm." *Procedia Manufacturing* 10 (2017): 1009-1019.

[40] Huang, Tingkai, Bingchan Li, Dongqin Shen, Jie Cao, and Bo Mao. "Analysis of the grain loss in harvest based on logistic regression." *Procedia computer science* 122 (2017): 698-705.

[41] Kazakevičiūtė, Agne, and Malini Olivo. "Point separation in logistic regression on Hilbert space-valued variables." *Statistics & Probability Letters* 128 (2017): 84-88.

[42] Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. "Heart disease prediction using machine learning and data mining technique." *Heart Disease* 7, no. 1 (2015): 129-137.

[43] Xing, Chao, Xin Geng, and Hui Xue. "Logistic boosting regression for label distribution learning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4489-4497. 2016.

[44] Siraj, Fadzilah, and Mansour Ali. *Mining enrollment data using descriptive and predictive approaches*. InTech—Open Access Company, 2011.

[45] Thiyagaraj, M., and G. Suseendran. "Research of Chronic Kidney Disease based on Data Mining Techniques." *International Journal of Recent Technology and Engineering (IJRTE). ISSN* (2019): 2277-3878.

[46] Elhoseny, Mohamed, K. Shankar, and J. Uthayakumar. "Intelligent diagnostic prediction and classification system for chronic kidney disease." *Scientific reports* 9, no. 1 (2019): 9583.

[47] Sobrinho, Alvaro, Andressa CM Da S. Queiroz, Leandro Dias Da Silva, Evandro De Barros Costa, Maria Eliete Pinheiro, and Angelo Perkusich. "Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques." *IEEE Access* 8 (2020): 25407-25419.

[48] Davenport, Thomas, and Ravi Kalakota. "The potential for artificial intelligence in healthcare." *Future healthcare journal* 6, no. 2 (2019): 94-98.

[49] Bhargava, Neeraj, Girja Sharma, Ritu Bhargava, and Manish Mathuria. "Decision tree analysis on j48 algorithm for data mining." *Proceedings of international journal of advanced research in computer science and software engineering* 3, no. 6 (2013).