Kidney Impairment Prediction Due to Diabetes Using Extended Ensemble Learning Machine Algorithm

[1]Deepa Devasenapathy, [2]Vidhya K, [3]Anna Alphy, [4]Finney Daniel Shadrach, [5]Jayaraj Velusamy and [6]Kathirvelu M

[1]Computing & Software Engineering, U.A. Whitaker College of Engineering, Florida Gulf Coast University, USA.

[2] Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India,

[3] Department of Computer Science and Engineering, SRM IST Delhi NCR campus Ghaziabad – 201204, India.

[4]Department of Electronics and Communication Engineering, KPR Institute of Engineering and Technology, India.

[5] Department of Electronics and Communication Engineering, Nehru Institute of Engineering and Technology, India.

[6] Department of Electronics and Communication Engineering, KPR Institute of Engineering and Technology, India.

[1]deepafgcu@gmail.com, [2]vidhyak@karunya.edu, [3]anna.urumbath@gmail.com, [4]finneydaniels@gmail.com, [5]jayarajmevlsi@gmail.com, [6]mkathirvelu77@gmail.com

Correspondence should be addressed to Vidhya K : vidhyak@karunya.edu.

Article Info

Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)

Doi: https://doi.org/10.53759/7669/jmc202303027

Received 10 January 2023; Revised from 16 April 2023; Accepted 20 May 2023.

Available online 05 July 2023.

Abstract – diabetes is the main cause for diabetic kidney disease (dkd), which affects the filtering units of kidneys slowly and stops it's function finally. This consequence is common for both genetic based (type 1) and lifestyle based (type 2) diabetes. However, type 2 diabetes plays a significant influence in increased urine albumin excretion, decreased glomerular filtration rate (gfr), or both. These causes failure of kidneys stage by stage. Herein, the implementation of extended ensemble learning machine algorithm (eelm) with improved elephant herd optimization (ieho) algorithm helps in identifying the severity stages of kidney damage. The data preprocessing and feature extraction process extracts three vital features such as period of diabetes (in year), gfr (glomerular filtration rate), albumin (creatinine ratio) for accurate prediction of kidney damage due to diabetes. Predicted result ensures the better outcome such as an accuracy of 98.869%, 97.899 % of precision ,97.993 % of recall and f-measure of 96.432 % as a result.

Keywords – Kidney Disease, Ensemble Learning, Diabetes, Elephant Herd Optimization.

I. INTRODUCTION

Kidney Disease is one of the major complications caused by diabetes. It is accounted for increased mortality among the population of Type 2 diabetes. Gheith, Osama et al .[1] have surveyed the prevalence and risk factors related to DKD. Microalbuminuria is more common in younger people, but lower GFR is more common in older people with DKD. El-Houssainy et al. [2] used machine learning approaches to detect and forecast the severity stages of chronic renal disease in the early stages. The classification model classifies the phases of CKD based on serum creatinine, blood urea, albumin, age, hemoglobin, and hypertension. Nicholas YQ Tan et al. [3], have analyzed the impact of sleep duration either short or long over the complications of Diabetic Kidney Disease. Yaeni Kim et al. [4], have addressed the complexity and Heterogeneity of diabetic nephropathy and suggested effective therapeutic agents for the management of diabetic nephropathy. Chan, L., Nadkarni et al. [5], have created machine learning based predictive model and tested against the components related to Diabetic Kidney Disease. The diagnosis of kidney infection is based on the patient's age, BMI, and estimated low Glomerular Filtration Rate (GFR), albuminuria, and creatinine, glucose, and hemoglobin concentrations (HbA1c).

MacIsaac RJ et al. [6] have clearly identified that the clinical trials of incretin-modulating drugs have indicated that they can lower albuminuria and possibly halt the rate of GFR of diabetic patients. Nayak et al. [7], applied EHO algorithm for

predicting the prevalence of cancer cells. They identified that classification techniques functions faster with feature selection and get slow during the absence of feature selection process. Velliangiri et al. [8], applied EHO for the detecting security attacks in cloud environment. The Algorithm is combined with fuzzy techniques for rules learning. The performance of the algorithm is evaluated using continuous simulations of computers and is compared with various state of art techniques. El Asnaoui et al. [9], applied the single and ensemble learning models for the pneumonia disease classification. The results obtained are compared with singe and combined form of MobileNet and ResNet 50 models. The performance metrics followed are accuracy, sensitivity, precision, recall F1-Score.

The ensemble model is elevated s a best performing model in pneumonia disease classification. Pérez, E., et al. [10], presented a melanoma detection convolutional neural network architecture based on ensemble learning and genetic algorithms. So an accurate prediction of affected level of DKD is an important life strengthening factor for diabetic patients. To acquire this, proposed method involves Improved Elephant Herd Optimization (IEHO) algorithm for feature extraction and Extended Ensemble Learning classifier for the classification of DKD levels.

II. RELATED WORK

Elshaarawy et al. [11] established an irrational and quick return to the genesis. Because of the balanced management clan updating operator and separating operator, the EEHO algorithm is more exploitative than the EHO algorithm. Wei Li et al. [12] developed an Improved Elephant Herd Optimization algorithm to enhance parameter control and selection, convergence speed, and efficiency of optimal solutions. An Ning & Ding et al. [13], implemented deep ensemble learning for Alzheimer's disease classification and obtained 4% better performance results over six ensemble algorithms. Gupta, A et al. [14] developed and tested an ensemble based for identifying Covid 19 related health issues which results good performance. Ibomoiye Domor Mienye et al. [15], created a model for predicting heart disease risk, where multiple CART models are combined into a homogenous ensemble model. ROCC is used to validate the accuracy of the suggested ensemble learning approach. Prasad et al. [16] employed machine learning techniques to assess kidney illness prediction. Naive Bayes, random forest, decision table, and J48 algorithms were tested, and their performance was assured for better detection of kidney problems caused by diabetes. Olayinka et al. [17], devised an ensemble approach to the diagnosis of chronic renal disease. The ensemble approaches such as Bagging and Random Subspace methods have effectively diagnosed the chronic kidney diseases.

Dong, Z et al. [18], created an ensemble model. The model predicted that DKD was more likely to occur in older T2DM patients with high homocysteine (Hcy), bad glycemic control, low serum albumin (ALB), low estimated glomerular filtration rate (eGFR), and high bicarbonate over the following three years. The ensemble model outperforms. Ghelichi-Ghojogh et al. [19], analyzed the links between CKD and a variety of behavioral and health-related factors in Iranian patients using logistic regression algorithm. The factors such as low birth weight, diabetes, chemotherapy are identified as most relevant causes of CKD. Xu et al. [20], applied random forest algorithm to predict diabetic kidney disease and obtained 89.831% of accuracy. The analysis involved totally 29 indicative markers including Microalbuminuria (ALB) and albumin-to-creatinine ratio etc. Based on the performance result, the confined the random forest algorithm is more suitable for clinical prediction of kidney diseases caused due to poor maintenance of diabetes.

Kandasamy Vidhya et al. [21], analyzed the possible complications of diabetes based on the habitual nature of patients. A Deep Belief Network (DBN) model constructed for disease prediction identifies the diabetes related risks depends on the day-to-day activities of patients. Ilyas et al[22] .'s model, which can successfully and sustainably identify all CKD phases, was developed using the Random Forest and J48 algorithms and the J48 algorithm is identified as the better one.

Kandasamy Vidhya et al. [23], implemented Modified Adaptive Neuro Fuzzy Inference System (MANFIS) to analyze the diseases prevalent in the society. Based on the multi-variate combinations of symptoms possible diseases are predicted.

Gazi et al. [24], performed a comparative study on CKD prediction using various algorithms and the LR algorithm is recognized as a best performing one based on the metrics of precision and accuracy. Satish Kumar David et al. [25], experimented with WEKA machine tool to predict the diabetic kidney disease by applying different techniques. The framework's effectiveness is evaluated using a variety of criteria, and the decision tree algorithms are found to be the most effective at forecasting DKD. Lin, CC et al. [26], developed a risk prediction model for CKD with the patients suffering from diabetes. The risk variables for CKD were identified using the Cox proportional hazards regression model. Violeta et al. [27], applied machine learning techniques to identify the biomarkers for diabetic nephropathy (DN). Determined that the techniques for accurate prediction that perform best are random forest and logistic regression.

Dunkler et al. [28], determined the relative influence of predictors using two risk prediction models for the occurrence and development of CKD after 5.5 years. For diabetes type 2 patients, albuminuria and eGFR were the most important markers in predicting the onset and progression of early CKD. Two machine learning algorithms were trained by Allen A et al. [29] to predict the phases of DKD severity, and their results were compared to the CDC risk score. The models were evaluated using both an external dataset compiled from various sources and a hold-out test set. A new temporal-enhanced gradient boosting machine (GBM) model was created by Song X et al. [30] that dynamically updates and groups learners in response to new events inpatient's life with greatest calibration in both moderate and high-risk categories. Gao et al. [31], created a model for

predicting renal function deterioration in individuals with type 2 DKD on an individual basis (T2DKD) and identified nomogram and risk table, are clinical indicators for predicting renal function deterioration in T2DKD patients at the bedside.

Materials And Methods

Poor diabetes treatment over time may cause kidney blood vessel clusters that filter waste from the circulation to become damaged, raising blood pressure. Renal disease is aggravated by high blood pressure, which increases pressure in the kidney's delicate filtering process. As an impact of kidney damage, the mortality count increases. So, an effective kidney mutilation prediction system is in need for the wellbeing of the diabetic community.

Dataset Description

The Chronic Kidney Disease Dataset gathered from UCI Machine Learning Repository is used to estimate the severity of CKD. 400 instances have 26 properties, including 14 nominal qualities and 12 numerical attributes. The cases are categorized well in advance as having CKD or not. Along with the existing attributes, 'gender' attribute is added. The total number of 26 attributes contains the elements of clinical and physiological in nature. The attribute information of the given dataset is given in **Table 1** below.

Table 1. Attributes Of CKD Dataset

S.No.	Name of the Attribute and its Specification		
1	Age(numerical) - age in years	14	Potassium(numerical) pot in mEq/L
2	Blood Pressure(numerical)- bp in mm/Hg	15	Hemoglobin(numerical) hemo in gms
3	Specific Gravity(nominal)- sg - (1.005,1.010,1.015,1.020,1.025)	16	Packed Cell Volume(numerical)
4	Albumin(nominal)- al - (0,1,2,3,4,5)	17	White Blood Cell Count (numerical) wc in cells/cumm
5	Sugar(nominal)- su - (0,1,2,3,4,5)	18	Red Blood Cell Count(numerical) rc in millions/cmm
6	Red Blood Cells(nominal) -rbc - (normal,abnormal)	19	Hypertension(nominal) htn - (yes,no)
7	Pus Cell (nominal)- pc - (normal,abnormal)	20	Diabetes Mellitus(nominal) dm - (yes,no)
8	Pus Cell clumps(nominal)- pcc - (present,notpresent)	21	Coronary Artery Disease(nominal) cad - (yes,no)
9	Bacteria(nominal)- ba - (present,notpresent)	22	Appetite(nominal) appet - (good,poor)
10	Blood Glucose Random(numerical) - bgr in mgs/dl	23	Pedal Edema(nominal) pe - (yes,no)
11	Blood Urea(numerical)- bu in mgs/dl	24	Anemia(nominal) ane - (yes,no)
12	Serum Creatinine(numerical)- sc in mgs/dl	25	Class (nominal) class - (ckd,notckd)
13	Sodium(numerical)- sod in mEq/L	26	Gender (Male/ Female)

Methodology Workflow

An Extended Ensemble Learning Machine (EELM) based machine learning approach is proposed to predict the infected stage of kidneys **Fig 1** depicts the suggested prediction model's design. At first, the data is collected from the UCI data repository and data pre-processing is done by performing missing value imputation, and data transformation. Then the feature selection is done by Improved Elephant Herd Optimization (IEHO) algorithm. Finally, the EELM algorithm classifies the infected levels of kidneys as NoDKD (No Diabetic Kidney Disease), Mild, Moderate, Severe and End-Stage-Renal Failure.

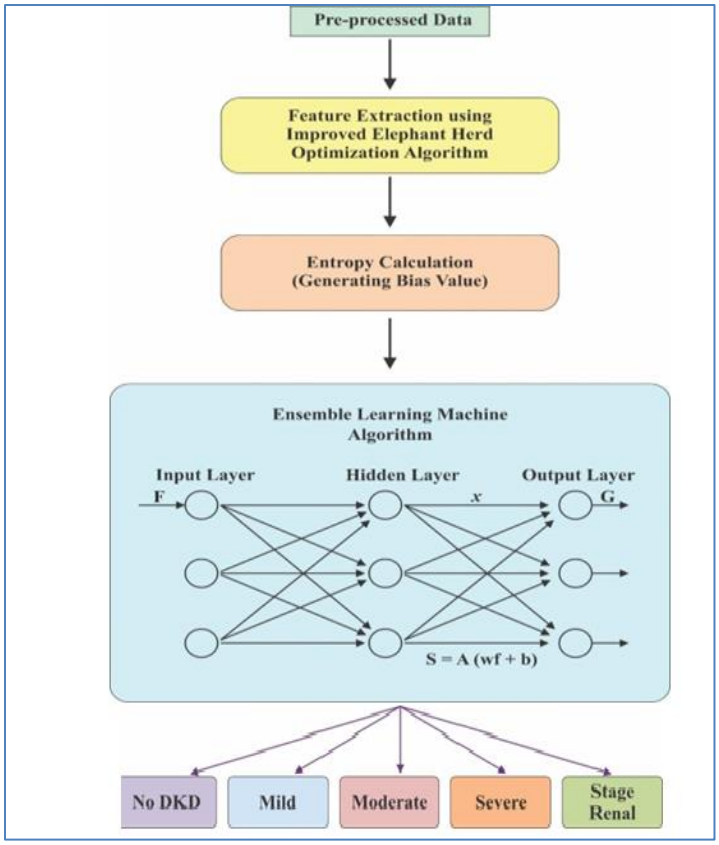


Fig 1. Architecture Of EELM For DKD Damage Level Prediction

Pre-Processing Of Diabetic Data

Data pre-processing is a method used for converting noisy and irrelevant data to clean one suitable for further utilization in analysis and prediction. It is an important process in data handling which lessens the dimensionality of data and helps to achieve better result. Data preprocessing is necessary before model development in order to remove a dataset's undesired noise and outliers that could cause the model to deviate from the intended training set. The effectiveness of the model is addressed at this step. The processing is done by implementing the techniques missing value imputation using mode. The missing values are filled by mode calculation.

Feature Extraction

The process of data reduction by removing extraneous data is closely related to feature extraction. The system reduces the dimensionality and simplifies the utilization which in turn minimizes the training time and improves accuracy. The IEHO algorithm extracts features in an optimized way.

	AGE	SC	GFR	SUGAR	AL	CLASS	CLASS TYPE	GENDER
0	48.0	1.2	75.0	0.0	1.0	ckd	2.0	f
1	17.0	0.8	110.0	0.0	4.0	ckd	1.0	m
2	62.0	1.8	43.0	3.0	2.0	ckd	3.0	m
3	48.0	3.8	19.0	0.0	4.0	ckd	4.0	f
4	51.0	1.4	46.0	0.0	2.0	ckd	3.0	m

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   AGE         391 non-null    float64
 1   SC          384 non-null    object
 2   GFR         380 non-null    float64
 3   SUGAR       351 non-null    float64
 4   AL          354 non-null    float64
 5   CLASS       400 non-null    object
 6   CLASS TYPE  388 non-null    float64
 7   GENDER      400 non-null    object
dtypes: float64(5), object(3)
memory usage: 25.1+ KB
```

Fig 2 . Significant Features

Fig 2 shows the attributes which are most related to kidney disease, such as patient's Age, Serum Creatinine (SC), Computed value of Glomerular Filtration Rate (GFR), Sugar Level, Albumin Urea (AU), Class (chronic kidney disease (CKD)/Non-CKD) and Class type (Infected level).

Improved Elephant Herd Optimization (Ieho) Algorithm

EHO is a global optimization technique that is modeled based on elephant behavior in nature. It does not make reference to previous measurements for present data processing. It varies from traditional meta-heuristic algorithms in this sense. If the features extracted from the previous model are fully exploited and used in other optimization process, the performance may be improved significantly. So, in the case of current implementation the EHO algorithm extracts highly significant three features from the previous iterations and is well enhanced by performing crossover and mutation. This procedure greatly increases the optimization effects. So, the proposed method for EHO is termed as Improved EHO (IEHO). Elephants' herding behaviors sparked the development of this algorithm. Elephants are sociable creatures in general having a composite social organization consisting of multiple clans (groups or networks) led by a matriarch. A clan is made up of one or more mother elephants and their calves. Male elephants like to live alone and will quit the clan as they grow older. Female elephants prefer to live in domestic clusters. Clan characteristics suggest exploitation, whereas abandoning elephants suggests population exploration. The features of elephants are evaluated using two key operators: clan updating and clan separating, which are used to improve clustering of Diabetic Kidney Disease-related attributes. The elephant population is represented by features such as Age, Blood pressure, Blood glucose Random, Hemoglobin, Diabetes Mellitus, Coronary Artery Disease, Anemia, Sugar, RBC count, Hypertension, Appetite, Pedal edoema, Packed cell volume, Pus cell, Pus cell clumps, Bacteria, and specific gravity.

The algorithm is originated by optimal parameter initialization and fitness function which evaluates the fitness of suitable features for the prediction of kidney damage level. The fitness function considers the solution obtained as candidate solution with respect to the problem of extracting related features. There is 'J' sub-KDPI networks for the whole set of attributes. Each network is led by a single dominating attribute that reflects the best dynamic KDPI solution. Each generation, attribute 'e' of dynamic $KDPIb_m$ goes closer to pbest, b_m, the dynamic $KDPIb_m$ with the best fitness. The accuracy values of the kidney disease complex parameters are used to calculate the dynamic KDPI's fitness. For new attributes e in dynamic KDPI b_m, the Equation (2) is referred for position updating.

$$h_{new,b_m,n} = h_{b_m,n} + \alpha\left(h_{best,b_m} - h_{b_m,v}\right) \times rand \tag{2}$$

where, $h_{new,b_m,n}$ denotes the new location of attributes n in clan m, $h_{b_m,n}$ signifies the position in previous generation, h_{best,b_m} denotes the best solution of clan b_m, $\alpha \in [0,1]$ is 'α' is the scale factor which identifies the best fitness and 'rand' represents the random number used to augment different population in later stages of algorithm. Equation(3) is used to update the optimum attribute position in clan.

$$h_{new,b_m} = \beta \times h_{center,b_m} \tag{3}$$

where, $\beta \in [0,1]$ specifies the next parameter of procedure followed and manages the functions of h_{center,b_m}, which is stated as in the Equation (4).

$$h_{center,b_m,d} = \frac{1}{n_{b_m}} \times \sum_{j=1}^{n_{e_m}} g_{b_m,j,d} \tag{4}$$

where, $1 \le d \le D$ specifies d^{th} element and D specifies overall space size and n_{b_m} specifies the number of attributes present in clan *m*. The best fitted attribute differs from their clan are reshuffled for modelling further.

In every clan a few attributes have the worst position value and are assigned a new position as shown in Equation 5.

$$h_{worst,b_m} = h_{min} + \left(h_{max} - h_{min} + 1\right) \times rand \tag{5}$$

where h_{min} denotes the search space least limit while h_{max} denotes the higher limit of search space. The $rand \in [0,1]$ is a random number chosen from a uniform distribution.

To make optimization more effective, crossover and mutation operations are performed when attribute positions are assessed. The 2-point cross-over is picked from among the several forms of crossovers. The parental qualities are given '2' points in the crossover selected. The genes in between the two locations are swapped between the parental and child traits, resulting in the child's attributes. These points are assessed as follows:

$$x_1 = \frac{\left|h_{new,b_m}\right|}{3} \tag{6}$$

$$x_2 = x_1 + \frac{\left|h_{new,b_m}\right|}{2} \tag{7}$$

As shown in **Equations 6 and 7**, the parameter swapping happens as of each attribute with new population. The newly created network of attributes are arbitrary populated until better fitness is acquired. The IEHO algorithm is explained as follows. The algorithm selects features such as diabetes Mellitus (nominal), sugar stages and albumin (Creatinine) ratio. Then based on the level of Serum Creatinine and patient's age the Glomerular Filtration rate is calculated.

Table 2. IEHO Algorithm for Feature Optimization of DKD Dataset

Input **:** Set of features related to DKD
Output : optimized features- Diabetes Mellitus, Glomerular Filtration Rate (GFR), albumin (Creatinine) level

Begin
Initialize Max Gen(KDPI Network) and DKD-attribute Population size
Set Maximum Generation M_G
Population $X_i = \{X_1, X_2, ...X_n\}$
Calculate fitness for each population of DKD attributes
While $t > M_G$ do
Sort all the attributes according to their fitness
for all KDPI network b_m in the population do
for all attributes j in the KDPI network c_m do
Updates $h_{b_{m,k}}$ and generate $h_{new,b_{m,k}}$ by using
$h_{new,b_{m,k}} = h_{b_{m,k}} + h(h_{best b_m} - h_{b_{m,k}}) \times rand$
if $h_{b_{m,k}} = h_{best b_m}$ then
Update $h_{b_{m,k}}$ and generate $h_{new,b_{m,k}}$ by using
$h_{new,b_m} = \delta \times h_{center} b_m$
end if
end for
end for
for all KDPI network in the DKD-attribute population do
Replace the worst attribute in the KDPI network
end for
Evaluate the DKD-attribute population and calculate fitness
end While
Perform crossover and mutation
Update the solution
Return the best suitable DKD-attribute among the KDPI network
End

Classification Of Kidney Damage Level Using Extended Ensemble Learning Machine (EELM) Algorithm

There is only one hidden layered feed forward neural network used in the Ensemble Learning Machine technique that generates weights between the output and hidden layers using the least square method. The bias values are randomly generated based on the input entropy and is called Extended ELM (EELM). Instead of generating random bias values, the proposed method uses functions of entropy as the selection criteria of optimal bias values. The Ensemble learning algorithm is extended by adding the performance of IEHO algorithm for selecting the optimal features. In addition, the entropy calculation approach is combined with the ensemble algorithm. The architecture of the proposed EELM is shown in **Table 2** and the steps involved in EELM are explained below.

Step 1: Initially, refer the training sample $X, Y = \{x_i, y_i\}$ where $i = 1,2,...Q$, and there is an input feature $X = \lfloor x_{i1}, x_{i2}, ...x_{iQ} \rfloor$ and a desired matrix $Y = \lfloor y_{i1}, y_{i2}, ...y_{iQ} \rfloor$ that consists training samples, where the matrix X and the matrix Y can be expressed as follows,

$$X = \begin{bmatrix} x_{11} & x_{12} \ldots\ldots\ldots\ldots x_{1Q} \\ x_{21} & x_{22} \ldots\ldots\ldots\ldots\ldots x_{2Q} \\ . & \\ . & \\ . & \\ x_{n1} & x_{n2} \ldots\ldots\ldots\ldots\ldots x_{nQ} \end{bmatrix}$$

s (8)

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \ldots\ldots\ldots\ldots Y_{1Q} \\ Y_{21} & Y_{22} \ldots\ldots\ldots\ldots\ldots Y_{2Q} \\ . & \\ . & \\ . & \\ Y_{m1} & Y_{m2} \ldots\ldots\ldots\ldots\ldots Y_{mQ} \end{bmatrix}$$

(9)

In this the values m and n specify the matrix for input and output.

Step 2: Assigned weight between input layer and the hidden layer is as mentioned in the weight matrix is represented in equation 10.

$$W = \begin{bmatrix} w_{11} & w_{12} \ldots\ldots\ldots\ldots w_{1n} \\ w_{21} & w_{22} \ldots\ldots\ldots\ldots\ldots w_{2n} \\ . & \\ . & \\ . & \\ w_{d1} & w_{d2} \ldots\ldots\ldots\ldots\ldots w_{dn} \end{bmatrix}$$

(10)

Step 3: The weight between the hidden and output layers is stated as follows.

$$\chi = \begin{bmatrix} \chi_{11} & \chi_{12} \cdots\cdots\cdots \chi_{1m} \\ \chi_{21} & \chi_{22} \cdots\cdots\cdots\cdots \chi_{2m} \\ . & \\ . & \\ . & \\ \chi_{d1} & \chi_{d2} \cdots\cdots\cdots\cdots \chi_{dm} \end{bmatrix} \tag{11}$$

Step 4: Assign the bias values for the invisible layer neurons by implementing the entropy values of the input weight values as given below.

$$b_i = -\sum_{i=1}^{n} w_i \, log_2 (w_i) \tag{12}$$

Step 5: Choose the SoftMax activation function $A\ (f)$, which calculates the probability distribution of factors for classifying the level of kidney damage. The probability range of SoftMax function is true (1) and false (1) and the total probability would be one. The prediction of kidney disease prediction majorly involves the class with maximum probability. The resultant matrix *H* is given as below,

$$H = \lfloor h_1, h_2, ... h_Q \rfloor_{m \times Q} \tag{13}$$

Column vector of the output matrix H is as follows:

$$h_j = \begin{bmatrix} h_{1j} \\ h_{2j} \\ . \\ . \\ h_{mj} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{d} \chi_{i1} A(w_i f_j + b_i) \\ \sum_{i=1}^{d} \chi_{i2} A(w_i f_j + b_i) \\ . \\ . \\ \sum_{i=1}^{d} \chi_{im} A(w_i f_j + b_i) \\ \end{bmatrix} \tag{14}$$

where $j = 1,2.3....Q$

Step 6: By referring Equations (5.12) and (5.13), compute $R\chi$ as per equation (15).

$$R\chi = K' \tag{15}$$

where *K'* is the transpose of K and $R\chi$ is the weight difference. The weight matrix values χ are calculated using the least square approach to obtain a unique solution with minimal error.

$$\chi = R^{+} K^{'} \tag{16}$$

A regularization term χ is also included to increase the network's generalization ability and make the findings more stable. When there are less hidden layer neurons than training data, the condition can be represented as,

$$\chi = \left(\frac{I}{\lambda} + R^H R \right)^{-1} R^H R^{'} \tag{17}$$

When there are more hidden layer nodes than the training data, it can be stated as,

$$\chi = R^H \left(\frac{I}{\lambda} + RR^H \right)^{-1} R^{'} \tag{18}$$

A standard ELM with d hidden neurons and activation function $A(f)$ are mathematically modelled by,

$$\sum_{j=1}^{d} \chi_j \, f\left(w_j f_i + b_j\right) = G_j \quad , 1 \le i \le n \tag{19}$$

where, the activation function f_i is calculated with the weighted function with bias value generated by the generator function G for the i^{th} , j^{th} layers of ELM structure involved.

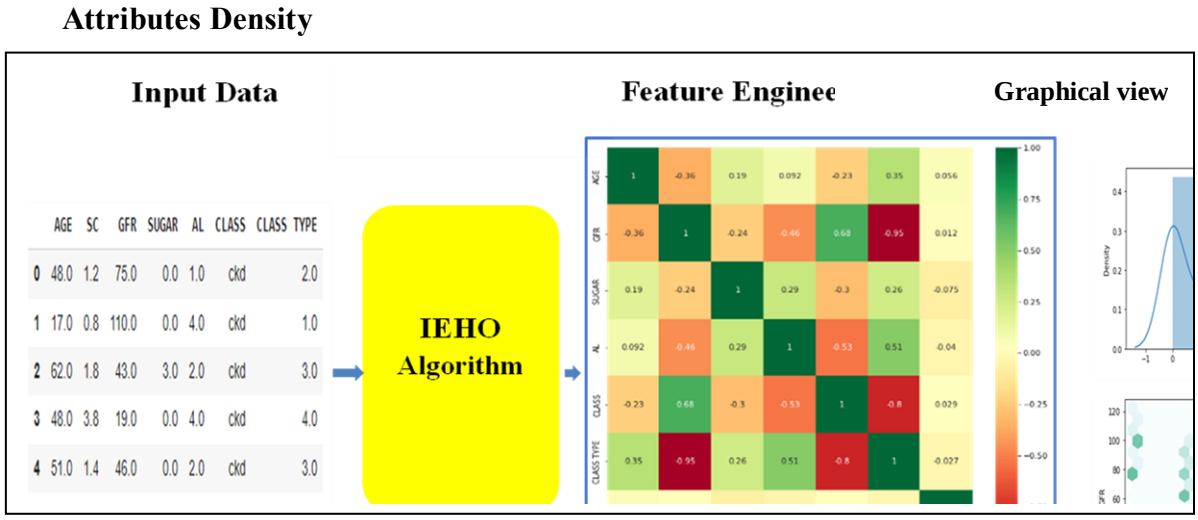


Fig 3. Visualization – Heat Map

Age, serum creatinine, glomerular filtration rate, sugar, albumin urea, class (CKD or Non-CKD), and class type (severity level) are having more of a direct impact on the development of diabetic kidney disease than other variables, according to the heat map in **Fig 3.**

Results and Discussion

Prediction of Kidney Damage Stages due to Diabetes

The proposed EELM algorithm, classifies the level of DKD into five stages such as NoDKD (0), mild (1), moderate (2), Severe (3), End-stage Renal (4) based on the period of diabetes, eGFR value and Albumin (Creatinine Ratio) as shown in **Table 3**. It is to be noted that the period of diabetes has a direct impact upon the functions of kidneys.

Table 3. Stages of Diabetic Kidney Damage Based on GFR And Albumin (Creatinine Ratio)

Diabetes Stages (Year)	eGFR (mL/min)	Albumin (Creatinine Ratio (mg/mmol))	Level of Kidney Damage (Stages)	Stage
2	>90	0-3(0)	NoDKD (Normal)	0
1	60-89	4-9(1)	Mild	1
5	30-59	10-25(2)	Moderate	2
4	15-29	26-30(3)	Severe	3
2	<15	>30(4)	End-Stage Renal Failure	4

The long persistence of diabetes can harm the blood vessels inside the kidneys which automatically degrade the filtration capacity of the kidney. So the GFR starts decreasing. Due to the reduced filtration rate, the albumin urea starts increasing and it is drained through filters. The level of kidney damage is classified into five stages as level 0 indicates that there is no damage of kidney due to diabetes, level 1 indicates that there is no kidney disease due to diabetes, level 2 is the moderate level of damage, level 3 is the indication of severe damage of kidney and finally level 4 indicates that the patient has gone to end stage of renal failure where the kidney completely stops its operation. Thus, the EELM algorithm effectively identifies the levels of kidney damage.

Performance Analysis Of EELM Algorithm

The performance of the IEHO method is evaluated using a confusion matrix, which is a tabular representation of the ability of a classification technique over test data in which the right and incorrect predictions are clearly shown. True Positive (TP) indicates that the expected component is positive or true, whereas False Positive (FP) indicates that the result is positive but is really false. False Negative (FN) is a false prediction that labels positive as negative. Negative values are correctly predicted as negative in True Negative (TN). The confusion matrix is constructed based on these interpretations, and the corresponding error rate, accuracy, and precision are employed.

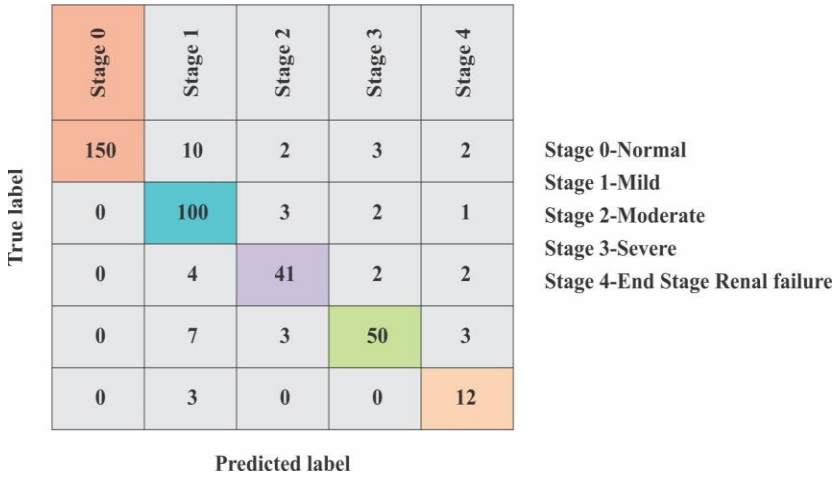


Fig 4. Confusion Matrix

The confusion matrix used to assess the effectiveness of the EELM Classifier is shown in **Fig 4**. The error rate is computed as 0.117, according to the confusion matrix. The accuracy, precision, recall, and F-measure values are calculated and the performance is compared with SVM, RBF, MLP and ELM algorithms.

Error Rate: The ratio between the sum of all the incorrect prediction and the sum of all the positive and the negative values are known as error rate. It is denoted below.

$$\text{Error rate} = \frac{(TP+TN)}{P+N} \tag{20}$$

Accuracy (ACC): It is defined as the methodical errors which measure the arithmetical biases caused between the true and the anticipated value.

$$\text{ACC} = \frac{(TP+TN)}{P+N} \tag{21}$$

Precision (P): Precision is the calculated as, among all the available positive classes, the number of classes predicted correctly as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{22}$$

Recall (R): It is the ratio between all the available classes and the number of classes predicted correctly as positive.

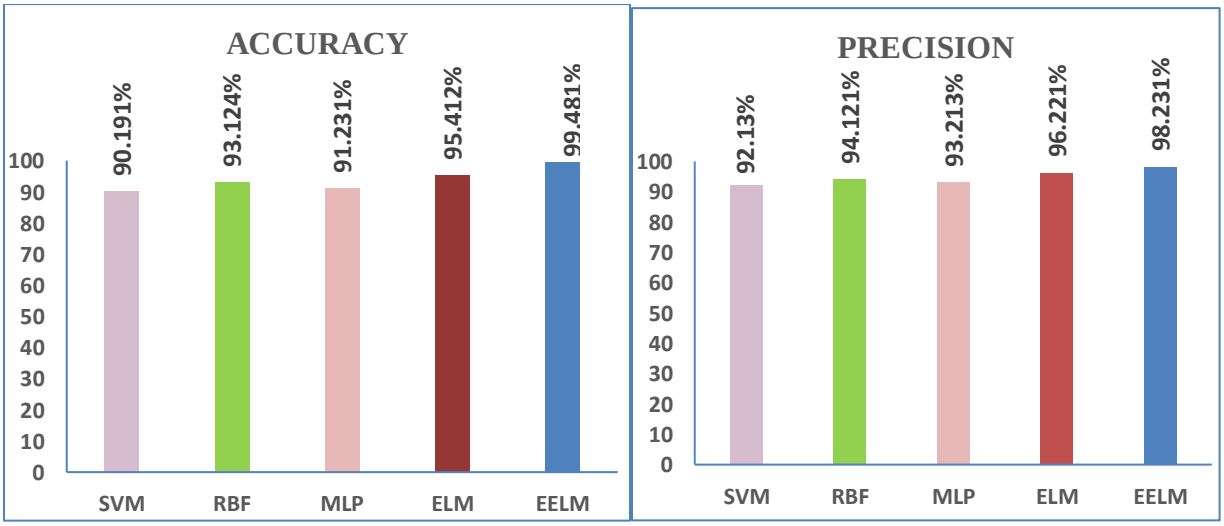
$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{23}$$

F1-Measure: It is calculated from precision and recall as shown below.

$$F1 - \text{Measure} = \frac{2\text{TP}}{(2\text{TP+FP+FN})} \tag{24}$$

The confusion matrix shown, conveys that the proposed EELM algorithm classifies the stage 0 accurately, which is of 150 patients, who have not been affected by the DKD and are normal, but the prediction in stage 1 has 24 false predictions, stage 2 has 8 false predictions, stage 3 has 7 false predictions and stage 4 has 8 false prediction values. So, the error rate of EELM is 0.117.

Performance of implemented EELM classification algorithm is measured with a precision of 99.481%, precision of 98.231%, recall of 98.953% and F-Measure of 98.582%. It is compared with the performance of other algorithms such as SVM, RBF, MLP and ELM techniques for classifying the affected stage of DKD as shown in **Fig 5**. Similarly, the visual inspection of the validation performance of EELM algorithm is shown in **Fig 6**. The performance of proposed EELM algorithm during validation is comparatively high. It obtains an accuracy of 98.860%, precision of 97.899%, recall of 97.993% and F-Measure of 97.899%.



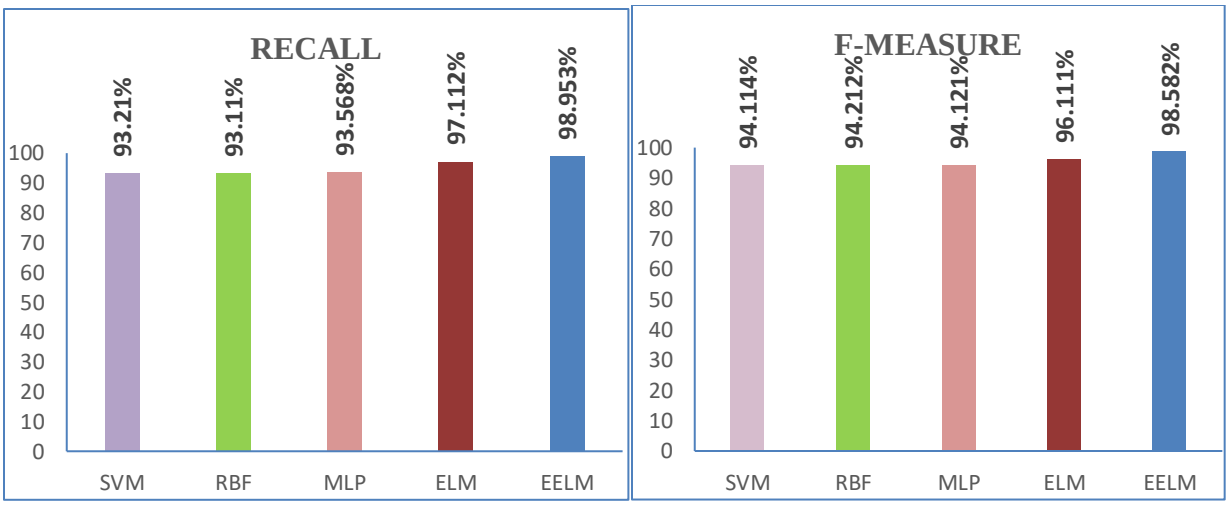


Fig 5. Comparison of Accuracy, Precision, Recall and F-Measure during Training

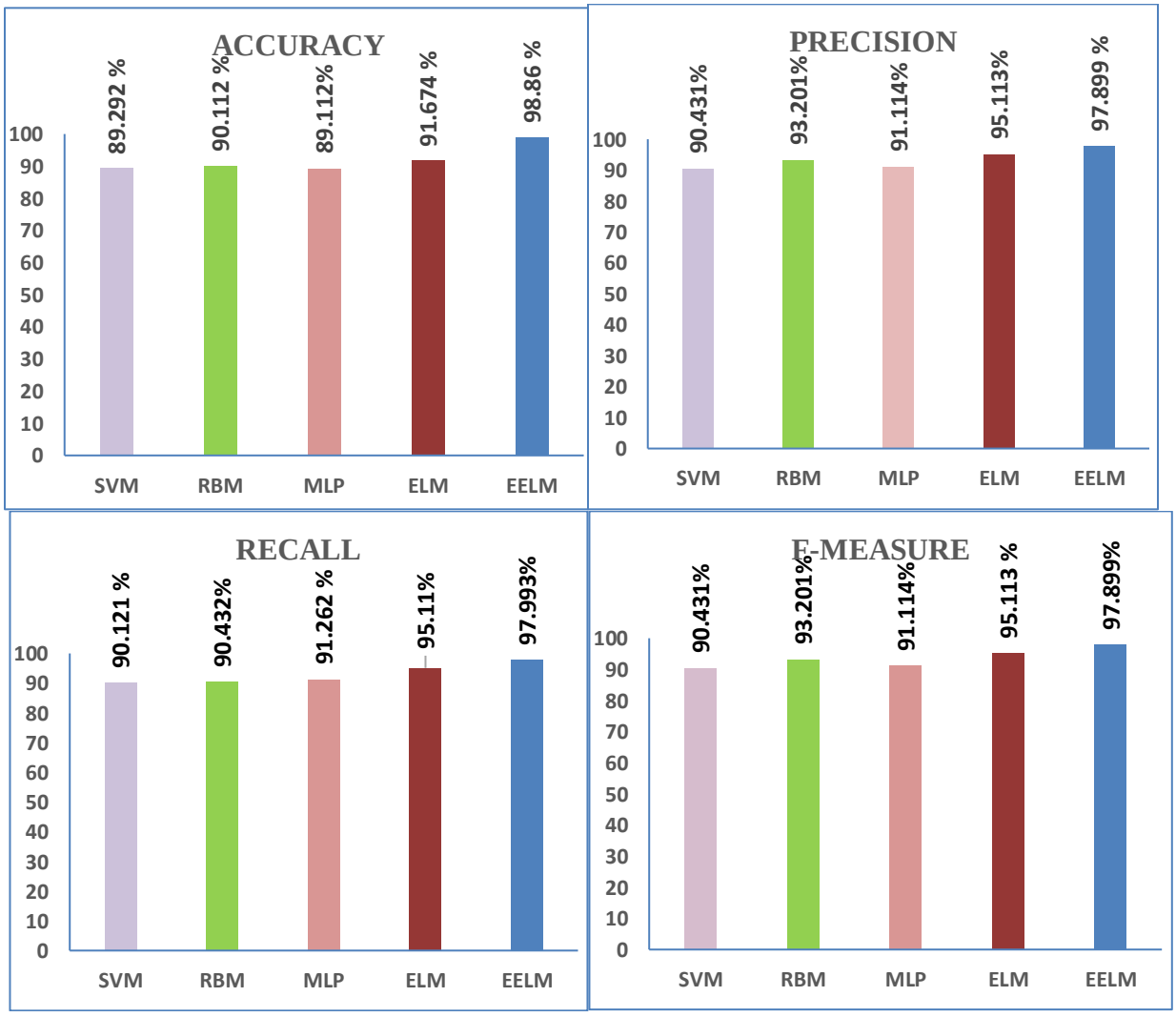


Fig 6. Comparison Of Accuracy, Precision, Recall And F-Measure During Validation

III.CONCLUSION

Thus the proposed EELM algorithm for kidney disease prediction determines the phases of kidney damage caused by diabetes automatically. This method classifies kidney damage into four categories: no DKD, mild, moderate, severe, and end-stage renal disease. The IEHO algorithm is used to choose a feature from the pre-processed data. For modifying the weights on the activation function, the IEHO and entropy-based bias value creation supports an optimum feature extraction. Furthermore, accuracy, precision, recall, and F-measure are used to assess the classification performance during both training and validation. When comparing the performance of the proposed system to that of individual classifiers, it is clear that extended ensemble learning classifiers outperform individual classifiers.

Availability Of Data And Materials

The dataset analyzed in this experiment was taken from the UCI repository where the CKD dataset is publicly available. The users can download the datasets from the link shared below. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics Approval and Consent to Participate

The research has consent for Ethical Approval and Consent to participate.

Competing Interests

There are no competing interests.

References

[1] Gheith, Osama & Othman, Nashwa & Nampoory, Narayanan & Halim, Medhat & Al-Otaibi, Torki., "Diabetic Kidney Disease Prevalence and Risk Factors". Journal of Nephropharmacology-Vol 5. pp 49–56, 2015

[2] El-Houssainy, Radya, A & Ayman S Anwar, "Prediction of kidney disease stages using data mining algorithms", Informatics in Medicine Unlocked, Elsevier, 2019

[3] Nicholas YQ Tan, Joel Chan, Ching-Yu Cheng, Tien Yin Wong & Charumathi Sabanayagam, "Sleep Duration and Diabetic Kidney Disease", Frontiers in Endocrinology, vol. 9, pp. 808, 2019

[4] Yaeni Kim & Cheol Whee Park, "New therapeutic agents in diabetic nephropathy", The Korean Journal of Internal Medicine, vol. 32, no. 1, pp. 11, 2017

[5] Chan, L., Nadkarni, G.N., Fleming, F. et al., "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease", Journal of Diabetologia vol.64, pp 1504–1515, 2021

[6] MacIsaac RJ, Thomas MC, "Effects of Diabetes Medications Targeting the Incretin System on the Kidney", Clin J Am Soc Nephrol, vol.13(2), pp 321-323, 2018

[7] Nayak, Monalisa & Das, Soumya & Urmila, Bhanja & Senapati, Manas Ranjan, " Elephant herding optimization technique based neural network for cancer prediction", Informatics in Medicine Unlocked, vol. 21,2020.

[8] S. Velliangiri, Hari Mohan Pandey, "Fuzzy-Taylor-elephant herd optimization inspired Deep Belief Network for DDoS attack detection and comparison with state-of-the-arts algorithms", Future Generation Computer Systems, vol 110, pp 80-90, 2020

[9] ElShaarawy, I.A., Houssein, E.H., Ismail, F.H. and Hassanien, A.E., "An Exploration-Enhanced Elephant Herding Optimization", Engineering Computations, vol. 36 No. 9, pp. 3029-3046, 2019

[10] Wei Li, Gai-Ge Wang, Amir H. Alavi,"Learning-based elephant herding optimization algorithm for solving numerical optimization problems", Knowledge-Based Systems, vol 195, 2020

[11] El Asnaoui, K. "Design Ensemble Deep Learning Model for Pneumonia Disease Classification", International Journal of Multimedia, Springer, pp 55 - 68 ,2021

[12] Pérez, E., Ventura, S, "An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis", Journal of Neural Compuing and Applications, 2021

[13] An, Ning & Ding, Huitong & Jiaoyun, Yang & Au, Rhoda & Ang, Ting Fang Alvin, "Deep Ensemble Learning for Alzheimer's Disease Classification",2019

[14] Gupta, A., Jain, V. & Singh, A, "Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications", Journal of New Generation Computers, 2021

[15] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, "An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk", Informatics in Medicine Unlocked, vol 20,2020

[16] Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N, "A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms", Journal of Computer science and Informatics, vol. 1, pp 101,2020
[17] Olayinka Ayodele Jongbo, Adebayo Olusola Adetunmbi, Roseline Bosede Ogunrinde, Bukola Badeji-Ajisafe, "Development of an Ensemble Approach to Chronic Kidney Disease Diagnosis , Journal of Scientific African Information Techniques, vol. 8,2020
[18] Dong, Z., Wang, Q., Ke, Y. et al., "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records", Journal of Translational Medicine, vol. 20, pp 143,2022
[19] Ghelichi-Ghojogh, M., Fararouei, M., Seif, M. et al., "Chronic kidney disease and its health-related factors: a case-control study", BMC Nephrol, vol 23, pp 24,2022
[20] Hongxia Xu, Yonghui Kong, and Shaofeng Tan, "Predictive Modeling of Diabetic Kidney Disease using Random Forest Algorithm along with Features Selection", International Symposium on Artificial Intelligence in Medical Science, 2020
[21] Kandasamy, Vidhya & Shanmugalakshmi, R, "Deep learning based big medical data analytic model for diabetes complication prediction", Journal of Ambient Intelligence and Humanized Computing, vol 11, 2020
[22] Ilyas, Hamida & Ali, Sajid & Ponum, Mahvish & Hasan, Osman & Mahmood, Muhammad & Iftikhar, Mehwish & Malik, Mubasher, "Chronic kidney disease diagnosis using decision tree algorithms", Journal of Nephrology, vol. 22.
[23] Kandasamy Vidhya & Shanmugalakshmi, R, "Modified adaptive neuro-fuzzy inference system (M-ANFIS) based multi-disease analysis of healthcare Big Data", Journal of Supercomputing, vol.11, 2020
[24] Gazi Mohammed Ifraz, Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis, Mohammad Monirujjaman Khan, "Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods", Journal of Computational and Mathematical Methods in Medicine, vol. 20, 2021
[25] Satish Kumar David, Mohamed Rafiullah, Khalid Siddiqui, "Comparison of Different Machine Learning Techniques to Predict Diabetic Kidney Disease", Journal of Healthcare Engineering, vol.10, 2022
[26] Lin, CC., Niu, M.J., Li, CI. et al., "Development and validation of a risk prediction model for chronic kidney disease among individuals with type 2 diabetes", vol. 11,2022
[27] Violeta Rodriguez-Romero, Richard F. Bergstrom, Brian S. Decker, Gezim Lahu, Majid Vakilynejad, Robert R. Bies, "Prediction of Nephropathy in Type 2 Diabetes: An Analysis of the ACCORD Trial Applying Machine Learning Techniques", Journal of Clinical and Translational Science", Volume12, Issue5, 2019
[28] Dunkler, Daniela and Gao, Peggy and Lee, Shun Fu and Heinze," Risk Prediction for Early CKD in Type 2 Diabetes", Clinical Journal of the American Society of Nephrology, vol.10,2015
[29] Allen A, Iqbal Z, Green-Saxena A, Hurtado M, Hoffman J, Mao Q, Das R., "Prediction of Diabetic Kidney Disease with Machine Learning Algorithms upon the initial diagnosis of Type 2 Diabetes Mellitus",2022
[30] Song X, Waitman LR, Yu AS, Robbins DC, Hu Y, Liu M, "Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study", JMIR Medical Informatics, 2020
[31] Gao YM, Feng ST, Yang Y, Li ZL, Wen Y, Wang B, Lv LL, Xing GL, Liu BC, "Development and External Validation of a Nomogram and a Risk Table for Prediction of Type 2 Diabetic Kidney Disease Progression Based on a Retrospective Cohort Study in China", Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, vol. 15, pp 799-811, 2022