Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Prediction of kidney disease stages using data mining algorithms

El-Houssainy A. Rady[a], Ayman S. Anwar[b,*]

[a] Institute of Statistical Studies & Research, Cairo University, Giza, Egypt
[b] Magrabi Hospitals & Centers, Cairo, Egypt

ARTICLE INFO

Keywords:
Prediction of kidney disease stages
Data mining techniques
Probabilistic neural networks
Multilayer perceptron
Support vector machine
Radial basis function

ABSTRACT

Early detection and characterization are considered to be critical factors in the management and control of chronic kidney disease. Herein, use of efficient data mining techniques is shown to reveal and extract hidden information from clinical and laboratory patient data, which can be helpful to assist physicians in maximizing accuracy for identification of disease severity stage. The results of applying Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Radial Basis Function (RBF) algorithms have been compared, and our findings show that the PNN algorithm provides better classification and prediction performance for determining severity stage in chronic kidney disease.

1. Introduction

A global health problem which is steadily growing is Chronic kidney disease (CKD). It is a chronic condition associated with increased morbidity and mortality, a high risk of many other diseases including cardiovascular disease, and high health care costs. Over two million people worldwide receive dialysis or kidney transplant treatment to stay alive, yet this number may represent only 10% of people who need treatment to live [9]. The majority of the 2 million people who receive treatment for kidney failure are in only five relatively wealthy countries, which represent 12% of the global population. By comparison, only 20% of the world's population is treated in about 100 developing countries, and they represent almost half the global population. Annually, more than one million people in 112 lower-income countries die from untreated kidney failure, due to the huge financial burden of dialysis or kidney transplantation treatment [9].

Thus, there is significant importance in the early detection, controlling, and managing of the disease. It is necessary to predict the progression of CKD with reasonable accuracy because of its dynamic and covert nature in the early stages, and patient heterogeneity. CKD is often described by severity stages. Clinical decisions are influenced by the stage, whether a patient is progressing, and the rate of progression. Also, defining the disease stage is quite crucial as it gives several indications that support the determination of required intervention and treatments.

Therefore, data mining can play a major role in extracting hidden data from the large patient medical and clinical dataset that physicians frequently collect from patients to obtain insights about the diagnostic information, and to implement precise treatment plans. Data mining can be defined as the process of extracting hidden data from a large dataset. Data mining techniques are applied and used widely in various contexts and fields. With data mining techniques we could predict, classify, filter and cluster data. The goal or prediction attribute refers to the algorithm processing of a training set containing a set of attributes and outcomes.

Machine learning algorithms have been used to predict and classify in the healthcare field. Yu et al. [17] have used the Support Vector Machine Algorithm to classify and predict diabetes and pre-diabetes patients, and the results show that SVM is useful to classify patients with common diseases. Similarly, Magnin et al. [19] have classified Alzheimer's disease by using a Support Vector Machine (SVM) to analyze whole-brain anatomical magnetic resonance imaging (MRI) for a set of patients, and the results shows that SVM is a promising approach for Alzheimer's disease early detection. Dessai et al. [18] have done heart disease prediction using the Probabilistic Neural Network Algorithm, Decision tree Algorithm, and Naïve Bayes Algorithm, and PRNN provides the best results compared with other algorithms for heart disease prediction. Cao et al. [20] have done prediction of HBV-induced liver cirrhosis using the Multilayered Perceptron (MLP) Algorithm and the results shows that the MLP classifier gives satisfactory prediction outputs for liver disease, mostly in HBV-related liver cirrhosis patients.

2. Materials & methods

Data Mining was utilized in our study because it is a process of identifying novel, potentially useful, valid and ultimately understandable patterns in data [4]. Supervised and unsupervised learning

[*] Corresponding author.
E-mail address: ayman.anwarr@gmail.com (A.S. Anwar).

https://doi.org/10.1016/j.imu.2019.100178
Received 9 December 2018; Received in revised form 11 March 2019; Accepted 1 April 2019
Available online 07 April 2019

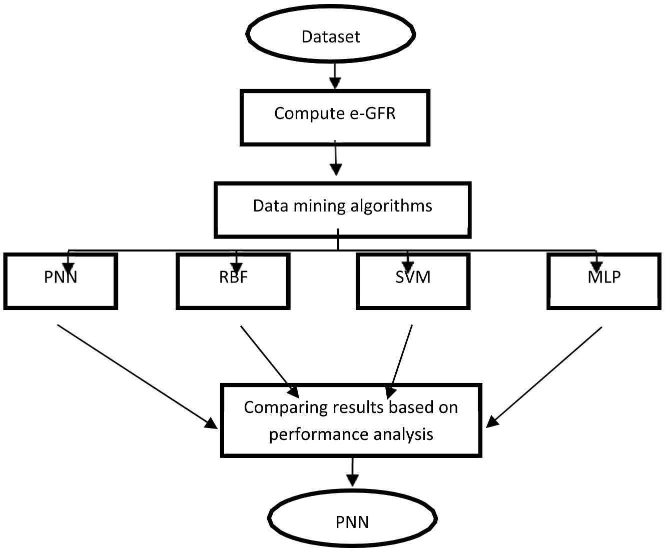

Fig. 1. Methodology workflow.

techniques are used for data mining classification. A “supervised” learning technique requires the building of a model based on previous performance analysis and is used in both medical and clinical research for classification, statistical regression and association rules [5]. On the other hand, the “unsupervised” learning technique is not guided by prior analysis and does not create a pre-analysis hypothesis. A model can be constructed based upon the results and is useful for clustering [6].

Three different types of the most commonly used artificial neural network algorithms and support vector machine algorithms have been used for this study, to determine which algorithm will give the best classification results, so as to identify the stage of chronic kidney disease, based on patient clinical and laboratory data. (see Fig. 1)

Machine learning techniques employ two phases to build the predictive/classification model as follows:

- A training phase that learns algorithmically how to build the model by using training datasets with expected outputs.
- A validation phase that estimates how well the model has been trained by using validation datasets without the expected outputs.

2.1. Probabilistic Neural Networks

Probabilistic Neural Networks (PNN) are a kind of Radial Basis Function neural network with a one pass learning algorithm and highly parallel structure. PNN was introduced by Donald F. Specht in 1990 as a memory-based network that provides estimates of categorical variables. The algorithm provides a smooth approximation of a target function, even with sparse data in a multidimensional space [16]. The advantages of PNN are fast learning and easy tuning. The PNN is composed of four layers: input, pattern (RBF kernel function), summation, and output, as shown in Fig. 2. Each neuron of the pattern layer uses a radial basis function as an activation function. This function is commonly taken to be Gaussian.

2.2. Multilayer Perceptron algorithm

The Multilayer Perceptron (MLP) is one of most important class of neural networks, consisting of an input layer, one or more hidden layers, and the output layer, as shown in Fig. 3. MLPs have been applied successfully to solve difficult and diverse problems, by training them in a supervised manner using a well-known algorithm i.e., the error back-propagation algorithm [3]. This algorithm is based on the error correction learning rule. As such, it may be viewed as a generalization of an adaptive filtering algorithm.

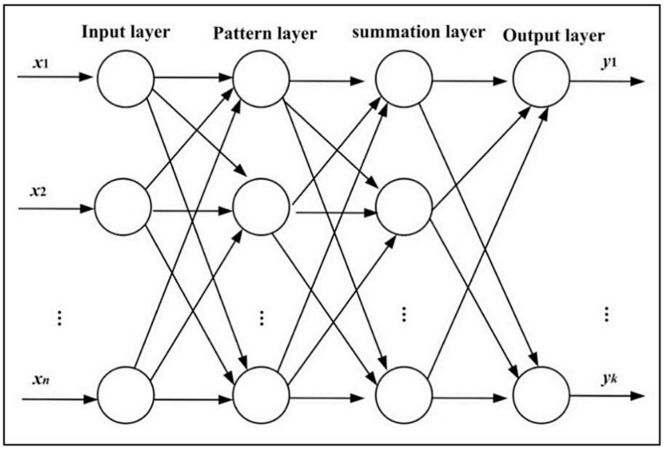

Fig. 2. Probabilistic neural networks (PNN) Layers.

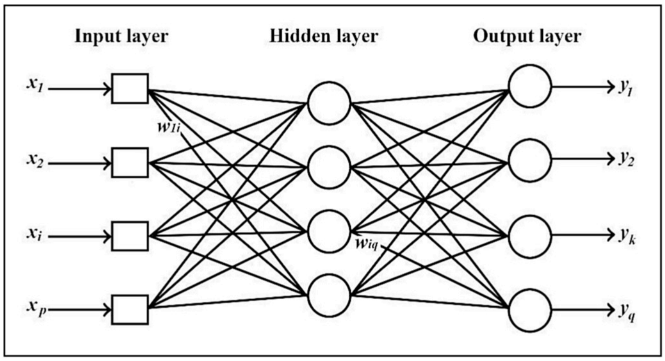

Fig. 3. Multilayer perceptron (MLP) layers.

2.3. Support vector machine algorithm

The SVM is a method for the classification of both linear and nonlinear data [7]. The SVM algorithm works as follows. It uses a nonlinear mapping to renovate the unique training data into a higher dimension. Surrounded by this new dimension, it examines the linear optimal separating hyperplane as shown in Fig. 4, i.e., a “decision boundary” sorting out the tuples of one class from another. With a suitable

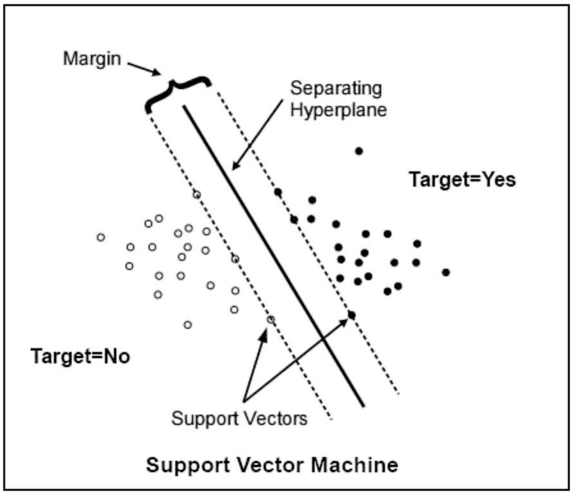

Fig. 4. Support Vector Machine (SVM) optimal hyperplane.

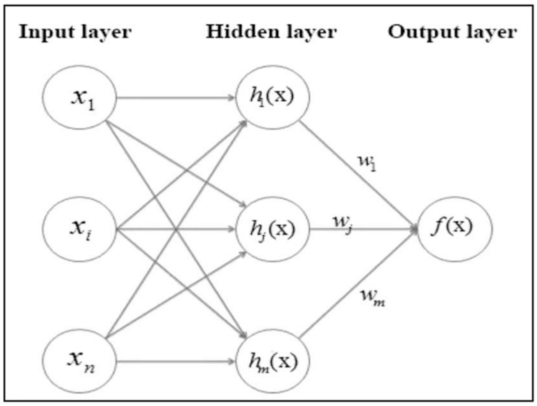

Fig. 5. Radial basis function (RBF) layers.

nonlinear mapping to a necessarily high dimension, data from two classes can always be separated by a hyperplane. The SVM finds the hyperplane using support vectors and margins [13]. Although the training time of even the fastest SVMs can be exceedingly slow, they are accurate, and exemplary in their ability to model complex nonlinear decision boundaries. They are much less prone to overfitting as compared with other methods. SVM initiates also provide a compact description of the learned model. SVMs can be used for prediction, along with classification. They have been applied to several areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests.

2.4. Radial basis function algorithm

The Radial Basis Function (RBF) is a neural network algorithm which requires less computing time for network training [10]. It consists of three layers: input layer, hidden layer, and output layer, as shown in Fig. 5 The nodes within each layer are fully connected to the previous layer [15]. The input variables in the input layer pass directly to the hidden layer without weights. The transfer functions of the hidden nodes are RBFs. The parameters associated with the RBFs are optimized during the network training. These parameter values are not necessarily the same throughout the network, nor are they directly related to or constrained by the actual training vectors. When the training vectors are assumed to be accurate, it is desirable to perform a smooth interpolation between them, then linear combinations of RBFs can be found which give no error at the training vectors. The method of fitting RBFs to data, for function approximation, is closely related to distance weighted regression.

3. Chronic kidney disease

CKD progression can be considered as a function of various parameters including underlying renal diseases, blood pressure, hypertension, proteinuria, and age. Early diagnosis of the CKD requires great attention among physicians, especially in determining the appropriate time to apply medical treatments and to control identified risk factors that reflect on the disease progression to End Stage Renal Disease (ESRD), such hypertension, proteinuria, and hyperphosphatemia.

3.1. Stages of chronic kidney disease

The stages of Chronic Kidney Disease (CKD) are mainly based on measured or estimated Glomerular Filtration Rate (eGFR). There are five stages, but kidney function is normal in Stage 1, and minimally reduced in Stage 2.

The KDOQI (Kidney Disease Outcomes Quality Initiative) stages of kidney disease are (see Table 1):

Definition of chronic: Labelling someone as having CKD requires two samples at least 90 days apart. Historical values can be used. The estimated Glomerular Filtration Rate (eGFR) depends on creatinine measurement, sex, race and age. One of the most accurate methods to calculate the eGFR is the Modification of Diet in Renal Disease (MDRD) [12].

$$\text{eGFR} = 186 \text{ x } (\text{Creatinine} / 88.4)^{-1.154} \text{ x } (\text{Age})^{-0.203} \text{ x } (0.742 \text{ if female}) \text{ x } (1.210 \text{ if black})$$

4. Results

The following analysis was performed using the DTREG Predictive Modeling System. The experimental comparison of the utilized algorithms was done based on the performance measures of classification accuracy and execution time. Model testing and validation was performed by a V-fold cross validation technique. Missing predictor variable values were replaced by medians during the analysis.

The dataset used in the analysis consisted of 361 CKD Indian patients and contained 25 variables (11 numerical, 14 categorical). Before starting the analysis, eGFR were calculated to identify the severity stage of the kidney disease for each patient by applying the eGFR formula described in section 3 on the used dataset. Dataset source is available on UCI machine learning repository.

4.1. Variables description

Follow in Table 2 is a description of variables used in the analysis, which contains the variable name, class, type, number of Missing rows, and categories, according to DTREG output.

4.2. Sensitivity and specificity

Sensitivity, Specificity and Accuracy percentage were employed to evaluate the performance of the utilized classification algorithms.

$$Sensitivity = \frac{TP}{TP\ + FN}$$

$$Specificity = \frac{TN}{TN\ + FP}$$

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

Table 1
CKD Stages according to GFR measurement value.

Stage	GFR	Description	Treatment stage
1	90+	Normal kidney function but urine findings or structural abnormalities or genetic trait point to kidney disease	Observation, control of blood pressure.
2	60–89	Mildly reduced kidney function, and other findings (as for stage 1) point to kidney disease	Observation, control of blood pressure and risk factors.
3A	45–59	Moderately reduced kidney function	Observation, control of blood pressure and risk factors.
3B	30–44		
4	15–29	Severely reduced kidney function	Planning for end stage renal failure.
5	< 15 or on dialysis	Very severe, or end stage kidney failure (sometimes call established renal failure)	Treatment choices.

Table 2
Variables description used in the analysis.

Ser	Variable	Class	Type	Missing rows	Categories
1	Age	Predictor	Numerical	0	66
2	Blood_Pressure	Predictor	Numerical	5	10
3	Specific_Gravity	Predictor	Categorical	7	5
4	Albumin	Predictor	Categorical	3	6
5	Sugar	Predictor	Categorical	4	6
6	Red_Blood_Cells	Predictor	Categorical	0	2
7	Pus_Cell	Predictor	Categorical	2	2
8	Pus_Cell_Clumps	Predictor	Categorical	4	2
9	Bacteria	Predictor	Categorical	4	2
10	Blood_Glucose_Random	Predictor	Numerical	5	143
11	Blood_Urea	Predictor	Numerical	3	116
12	Serum_Creatinine	Predictor	Numerical	0	83
13	Sodium	Predictor	Numerical	0	34
14	Potassium	Predictor	Numerical	7	39
15	Hemoglobin	Predictor	Numerical	8	111
16	Packed_Cell_Volume	Predictor	Numerical	10	41
17	White_Blood_Cell_Count	Predictor	Numerical	7	86
18	Red_Blood_Cell_Count	Predictor	Numerical	7	45
19	Hypertension_	Predictor	Categorical	2	2
20	Diabetes_Mellitus_	Predictor	Categorical	2	2
21	Coronary_Artery_Disease_	Predictor	Categorical	2	2
22	Sex_	Predictor	Categorical	1	2
23	Pedal_Edema_	Predictor	Categorical	1	2
24	Anemia_	Predictor	Categorical	1	2
25	CKD_STAGE_	Target	Categorical	0	5

Table 3
Summary of Algorithms classification outputs for classifying the CKD patients with stage 1 disease severity.

Sensitivity & Specificity – CKD Stage 1	PNN		SVM		RBF		MLP	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Total records	361	361	361	361	361	361	361	361
Accuracy	100.00%	99.72%	99.45%	91.14%	98.89%	95.84%	77.56%	77.29%
True positive (TP)	79 (21.88%)	79 (21.88%)	78 (21.61%)	57 (15.79%)	76 (21.05%)	72 (19.94%)	0 (0.00%)	11 (3.05%)
True negative (TN)	282 (78.12%)	281 (77.84%)	281 (77.84%)	272 (75.35%)	281 (77.84%)	274 (75.90%)	280 (77.56%)	268 (74.24%)
False positive (FP)	0 (0.00%)	1 (0.28%)	1 (0.28%)	10 (2.77%)	1 (0.28%)	8 (2.22%)	2 (0.55%)	14 (3.88%)
False negative (FN)	0 (0.00%)	0 (0.00%)	1 (0.28%)	22 (6.09%)	3 (0.83%)	7 (1.94%)	79 (21.88%)	68 (18.84%)
Sensitivity	100.00%	100.00%	98.73%	72.15%	96.20%	91.14%	0.00%	13.92%
Specificity	100.00%	99.65%	99.65%	96.45%	99.65%	97.16%	99.29%	95.04%
Geometric mean of sensitivity and specificity	100.00%	99.82%	99.19%	83.42%	97.91%	94.10%	0.00%	36.38%
Positive Predictive Value (PPV)	100.00%	98.75%	98.73%	85.07%	98.70%	90.00%	0.00%	44.00%
Negative Predictive Value (NPV)	100.00%	100.00%	99.65%	92.52%	98.94%	97.51%	77.99%	79.76%
Geometric mean of PPV and NPV	100.00%	99.37%	99.19%	88.72%	98.82%	93.68%	0.00%	59.24%
Precision	100.00%	98.75%	98.73%	85.07%	98.70%	90.00%	0.00%	44.00%
Recall	100.00%	100.00%	98.73%	72.15%	96.20%	91.14%	0.00%	13.92%
F-Measure	1	0.9937	0.9873	0.7808	0.9744	0.9057	0	0.2115

Table 4
Summary of Algorithms classification outputs for classifying the CKD patients with stage 2 disease severity.

Sensitivity & Specificity – CKD Stage 2	PNN		SVM		RBF		MLP	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Total records	361	361	361	361	361	361	361	361
Accuracy	100.00%	98.89%	99.45%	85.60%	93.63%	90.58%	72.85%	71.47%
True positive (TP)	81 (22.44%)	78 (21.61%)	80 (22.16%)	53 (14.68%)	80 (22.16%)	75 (20.78%)	75 (20.78%)	58 (16.07%)
True negative (TN)	280 (77.56%)	279 (77.29%)	279 (77.29%)	256 (70.91%)	258 (71.47%)	252 (69.81%)	188 (52.08%)	200 (55.40%)
False positive (FP)	0 (0.00%)	1 (0.28%)	1 (0.28%)	24 (6.65%)	22 (6.09%)	28 (7.76%)	92 (25.48%)	80 (22.16%)
False negative (FN)	0 (0.00%)	3 (0.83%)	1 (0.28%)	28 (7.76%)	1 (0.28%)	6 (1.66%)	6 (1.66%)	23 (6.37%)
Sensitivity	100.00%	96.30%	98.77%	65.43%	98.77%	92.59%	92.59%	71.60%
Specificity	100.00%	99.64%	99.64%	91.43%	92.14%	90.00%	67.14%	71.43%
Geometric mean of sensitivity and specificity	100.00%	97.96%	99.20%	77.35%	95.40%	91.29%	78.85%	71.52%
Positive Predictive Value (PPV)	100.00%	98.73%	98.77%	68.83%	78.43%	72.82%	44.91%	42.03%
Negative Predictive Value (NPV)	100.00%	98.94%	99.64%	90.14%	99.61%	97.67%	96.91%	89.69%
Geometric mean of PPV and NPV	100.00%	98.84%	99.20%	78.77%	88.39%	84.33%	65.97%	61.40%
Precision	100.00%	98.73%	98.77%	68.83%	78.43%	72.82%	44.91%	42.03%
Recall	100.00%	96.30%	98.77%	65.43%	98.77%	92.59%	92.59%	71.60%
F-Measure	1	0.975	0.9877	0.6709	0.8743	0.8152	0.6048	0.5297

Table 5
Summary of Algorithms classification outputs for classifying the CKD patients with stage 3 disease severity.

Sensitivity & Specificity - CKD Stage 3	PNN		SVM		RBF		MLP	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Total records	361	361	361	361	361	361	361	361
Accuracy	100.00%	96.95%	100.00%	73.68%	94.18%	92.52%	82.55%	77.29%
True positive (TP)	82 (22.71%)	81 (22.44%)	82 (22.71%)	48 (13.30%)	62 (17.17%)	61 (16.90%)	63 (17.45%)	56 (15.51%)
True negative (TN)	279 (77.29%)	269 (74.52%)	279 (77.29%)	218 (60.39%)	278 (77.01%)	273 (75.62%)	235 (65.10%)	223 (61.77%)
False positive (FP)	0 (0.00%)	10 (2.77%)	0 (0.00%)	61 (16.90%)	1 (0.28%)	6 (1.66%)	44 (12.19%)	56 (15.51%)
False negative (FN)	0 (0.00%)	1 (0.28%)	0 (0.00%)	34 (9.42%)	20 (5.54%)	21 (5.82%)	19 (5.26%)	26 (7.20%)
Sensitivity	100.00%	98.78%	100.00%	58.54%	75.61%	74.39%	76.83%	68.29%
Specificity	100.00%	96.42%	100.00%	78.14%	99.64%	97.85%	84.23%	79.93%
Geometric mean of sensitivity and specificity	100.00%	97.59%	100.00%	67.63%	86.80%	85.32%	80.44%	73.88%
Positive Predictive Value (PPV)	100.00%	89.01%	100.00%	44.04%	98.41%	91.04%	58.88%	50.00%
Negative Predictive Value (NPV)	100.00%	99.63%	100.00%	86.51%	93.29%	92.86%	92.52%	89.56%
Geometric mean of PPV and NPV	100.00%	94.17%	100.00%	61.72%	95.82%	91.95%	73.81%	66.92%
Precision	100.00%	89.01%	100.00%	44.04%	98.41%	91.04%	58.88%	50.00%
Recall	100.00%	98.78%	100.00%	58.54%	75.61%	74.39%	76.83%	68.29%
F-Measure	1	0.9364	1	0.5026	0.8552	0.8188	0.6667	0.5773

Table 6
Summary of Algorithms classification outputs for classifying the CKD patients with stage 4 disease severity.

Sensitivity & Specificity – CKD Stage 4	PNN		SVM		RBF		MLP	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Total records	361	361	361	361	361	361	361	361
Accuracy	100.00%	99.72%	100.00%	80.33%	99.45%	96.95%	86.43%	85.32%
True positive (TP)	57 (15.79%)	56 (15.51%)	57 (15.79%)	15 (4.16%)	56 (15.51%)	49 (13.57%)	12 (3.32%)	9 (2.49%)
True negative (TN)	304 (84.21%)	304 (84.21%)	304 (84.21%)	275 (76.18%)	303 (83.93%)	301 (83.38%)	300 (83.10%)	299 (82.83%)
False positive (FP)	0 (0.00%)	0 (0.00%)	0 (0.00%)	29 (8.03%)	1 (0.28%)	3 (0.83%)	4 (1.11%)	5 (1.39%)
False negative (FN)	0 (0.00%)	1 (0.28%)	0 (0.00%)	42 (11.63%)	1 (0.28%)	8 (2.22%)	45 (12.47%)	48 (13.30%)
Sensitivity	100.00%	98.25%	100.00%	26.32%	98.25%	85.96%	21.05%	15.79%
Specificity	100.00%	100.00%	100.00%	90.46%	99.67%	99.01%	98.68%	98.36%
Geometric mean of sensitivity and specificity	100.00%	99.12%	100.00%	48.79%	98.96%	92.26%	45.58%	39.41%
Positive Predictive Value (PPV)	100.00%	100.00%	100.00%	34.09%	98.25%	94.23%	75.00%	64.29%
Negative Predictive Value (NPV)	100.00%	99.67%	100.00%	86.75%	99.67%	97.41%	86.96%	86.17%
Geometric mean of PPV and NPV	100.00%	99.84%	100.00%	54.38%	98.96%	95.81%	80.76%	74.43%
Precision	100.00%	100.00%	100.00%	34.09%	98.25%	94.23%	75.00%	64.29%
Recall	100.00%	98.25%	100.00%	26.32%	98.25%	85.96%	21.05%	15.79%
F-Measure	1	0.9912	1	0.297	0.9825	0.8991	0.3288	0.2535

Table 7
Summary of Algorithms classification outputs for classifying the CKD patients with stage 5 disease severity.

Sensitivity & Specificity - CKD Stage 5	PNN		SVM		RBF		MLP	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Total records	361	361	361	361	361	361	361	361
Accuracy	100.00%	98.06%	100.00%	90.58%	100.00%	98.06%	95.84%	91.69%
True positive (TP)	62 (17.17%)	55 (15.24%)	62 (17.17%)	46 (12.74%)	62 (17.17%)	57 (15.79%)	58 (16.07%)	52 (14.40%)
True negative (TN)	299 (82.83%)	299 (82.83%)	299 (82.83%)	281 (77.84%)	299 (82.83%)	297 (82.27%)	288 (79.78%)	279 (77.29%)
False positive (FP)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (4.99%)	0 (0.00%)	2 (0.55%)	11 (3.05%)	20 (5.54%)
False negative (FN)	0 (0.00%)	7 (1.94%)	0 (0.00%)	16 (4.43%)	0 (0.00%)	5 (1.39%)	4 (1.11%)	10 (2.77%)
Sensitivity	100.00%	88.71%	100.00%	74.19%	100.00%	91.94%	93.55%	83.87%
Specificity	100.00%	100.00%	100.00%	93.98%	100.00%	99.33%	96.32%	93.31%
Geometric mean of sensitivity and specificity	100.00%	94.19%	100.00%	83.50%	100.00%	95.56%	94.92%	88.47%
Positive Predictive Value (PPV)	100.00%	100.00%	100.00%	71.88%	100.00%	96.61%	84.06%	72.22%
Negative Predictive Value (NPV)	100.00%	97.71%	100.00%	94.61%	100.00%	98.34%	98.63%	96.54%
Geometric mean of PPV and NPV	100.00%	98.85%	100.00%	82.46%	100.00%	97.47%	91.05%	83.50%
Precision	100.00%	100.00%	100.00%	71.88%	100.00%	96.61%	84.06%	72.22%
Recall	100.00%	88.71%	100.00%	74.19%	100.00%	91.94%	93.55%	83.87%
F-Measure	1	0.9402	1	0.7302	1	0.9421	0.8855	0.7761

Where:

- TP is Number of true positive classification cases
- FN is Number of false negative classification cases
- TN is Number of true negative classification cases
- FP is Number of false positive classification cases

Algorithm classification results are exhibited in Table (3) for patients with CKD stage 1 disease severity, and shows that PNN Algorithm gives the highest classification accuracy of 99.7%, Precision 98.7% and F-Measure 99.37% as compared with all other algorithm results.

Algorithm classification results are displayed in Table (4) for patients with CKD stage 2 disease severity, and shows that PNN algorithm

Table 8
Overall classification accuracy percentage and analysis execution time for all used algorithms.

Algorithm	Overall Accuracy	Total Execution Time
PNN	96.7%	0:00:12
SVM	60.7%	0:00:40
RBF	87%	2:29.6
MLP	51.5%	00:03.5

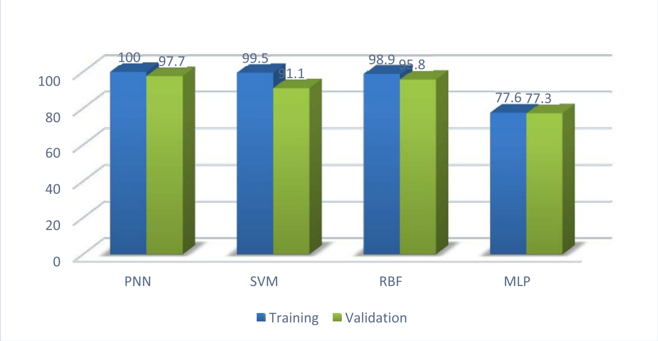

Fig. 6. CKD Stage 1 Classification accuracy % for all used algorithms.

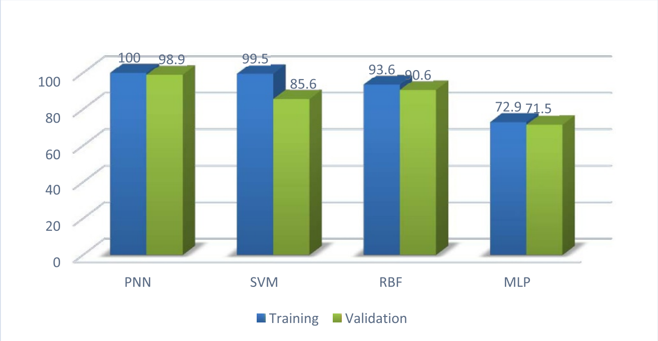

Fig. 7. CKD Stage 2 Classification accuracy % for all used algorithms.

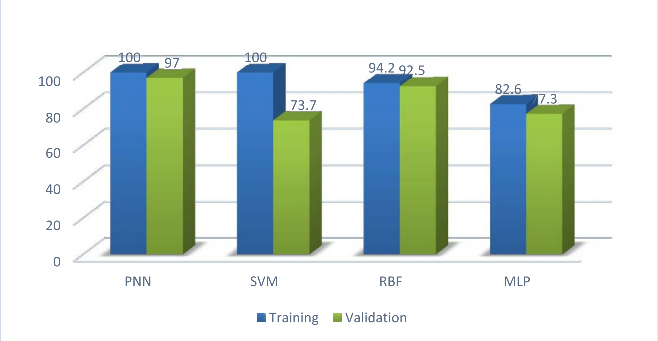

Fig. 8. CKD Stage 3 Classification accuracy % for all used algorithms.

provides a highest classification accuracy 98.9%, Precision 98.7% and F— Measure 97.5% as compared with all other algorithm results.

Algorithm classification results in Table 5 for the patients with CKD stage 3 disease severity shows that the PNN algorithm gives highest classification accuracy with percentage 96.9%, Precision 89% and F-Measure 93.6% as compared with all other algorithm results.

Algorithm classification results in Table 6 for patients with CKD stage 4 disease severity shows that the PNN algorithm gives the highest classification accuracy with a percentage 99.7%, Precision 100% and F-Measure 99.1% as compared with all other algorithm results.

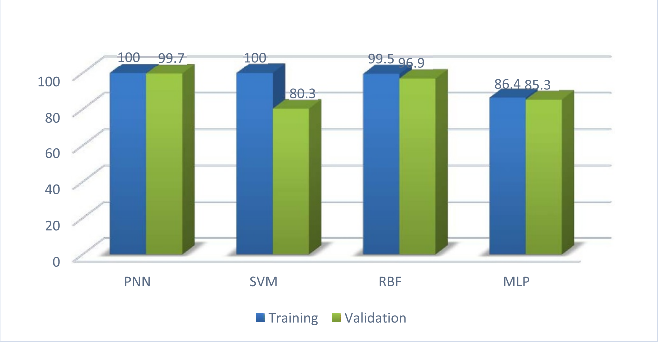

Fig. 9. CKD Stage 4 Classification accuracy % for all used algorithms.

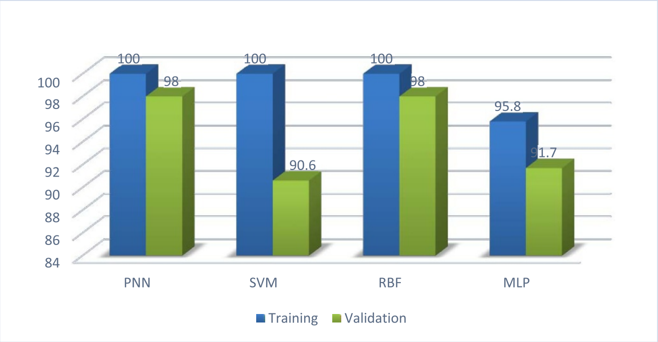

Fig. 10. CKD Stage 5 Classification accuracy % for all used algorithms.

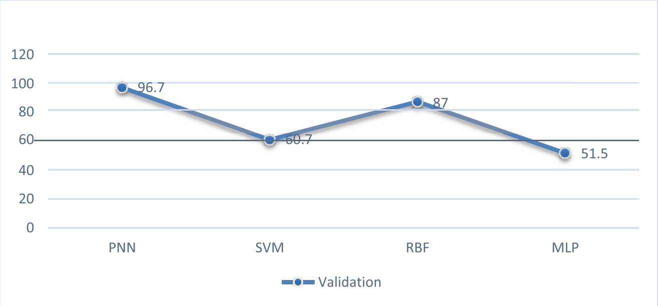

Fig. 11. Overall Classification accuracy % for all used algorithms.

Algorithm classification results in Table 7 for patients with CKD stage 5 disease severity shows that the PNN algorithm provides the highest classification accuracy with percentage 98%, Precision 100% and F-Measure 94% as compared with all other algorithm results.

5. Discussion

Our study suggests that the severity stages of chronic kidney disease

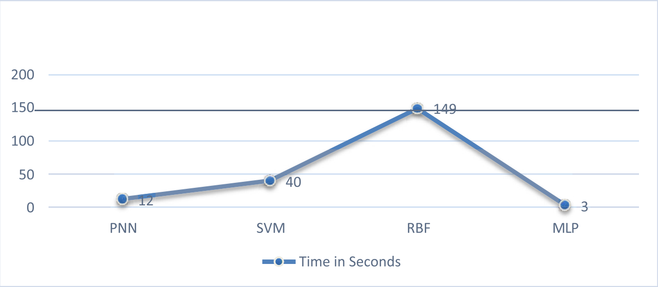

Fig. 12. Total analysis execution time in seconds for all used algorithms.

Table 9
Overall importance of predictive variables in building the classification model.

Predictive Variables	Importance %
Serum Creatinine	100.000
Blood Urea	38.455
Albumin	22.034
Age	20.394
Hemoglobin	10.359
Hypertension	9.256

can be accurately classified and predicted by using data mining techniques. The above-mentioned results from Tables 3–8 suggest that the Probabilistic Neural Networks and Radial Basis Function techniques are providing the most accurate classification, precision, and highest F-Measure, comparable with the Support Vector Machine and Multilayer Perceptron techniques. On the other hand, the Radial Basis Function technique requires more processing time than the Probabilistic Neural Network technique.

The Probabilistic Neural Network technique gives best classification results as compared with all other used techniques in classifying CKD stages (see Figs. 6-12), as follows:

o Accuracy percentage 99.7%, Precision 98.7% and F-Measure 99.37% in classifying Stage 1 CKD patients.
o Accuracy percentage 98.9%, Precision 98.7% and F-Measure 97.5% in classifying Stage 2 CKD patients.
o Accuracy percentage 96.9%, Precision 89% and F-Measure 93.6% in classifying Stage 3 CKD patients.
o Accuracy percentage 99.7%, Precision 100% and F-Measure 99.1% in classifying Stage 4 CKD patients.
o Accuracy percentage 98%, Precision 100% and F-Measure 94% in classifying Stage 5 CKD patients.

The results of Table 9 shows that the following predictor variables are the most important variables during construction of the classification model: Serum Creatinine (100%), Blood Urea (38.5%), Albumin (22%), Age (20%), Hemoglobin (10%) and Hypertension (9%).

The Probabilistic Neural Networks technique can be readily implemented for classifying the severity stages of chronic kidney disease patients.

6. Conclusion

Finally, as observed from Table 8, the Probabilistic Neural Networks algorithm gives the highest overall classification accuracy percentage of 96.7%, compared to other algorithms in classifying the stages of CKD patients. On the other hand, the Multilayer Perceptron requires a minimum execution time (3 s) whereas the Probabilistic Neural Network requires 12 s to finalize the analysis.

These algorithms have been compared with classification accuracy based on correctly classified stages of CKD patients, time taken to construct the model, and time taken to test the model. The Probabilistic Neural Networks algorithm yields a better classification accuracy and prediction performance to predict the stages of chronic kidney disease patients.

Significance Statement: The current study applied four data mining algorithms on a clinical/laboratory dataset consisting of 361 chronic kidney disease patients. The results of the addressed algorithms have been compared to define the most accurate algorithm results in classifying the severity stage of CKD. This study recommends that the Probabilistic Neural Networks algorithm is the best algorithm that can be used by physicians in order to eliminate diagnostic and treatment errors.

Conflicts of interest

No competing interest exists.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2019.100178.

References

[3] Ramchoun H, Amine M, Idrissi J, Ghanou Y, Ettaouil M. Multilayer Perceptron: Architecture optimization and training. IJIMAI 2016;4(1):26–30.
[4] Dhamodharan S. Liver disease prediction using bayesian classification. 4th national conference on advance computing. Application Technologies; 2014, May.
[5] Joshi J, Doshi R, Patel J. Diagnosis and prognosis breast cancer using classification rules. Int J Eng Res Gen Sci 2014;2(6):315–23.
[6] Solanki AV. Data mining techniques using WEKA classification for sickle cell disease. Int J Comput Sci Inf Technol 2014;5(4):5857–60.
[7] Aljahdali S, Hussain SN. Comparative prediction performance with support vector machine and random forest classification techniques. Int J Comput Appl 2013;69(11).
[9] Couser WG, Remuzzi G, Mendis S, Tonelli M. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. Kidney Int 2011;80(12):1258–70.
[10] Schaback R. A practical guide to radial basis functions. Electronic Resource 2007;11.
[12] Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, et al. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Ann Intern Med 2003;139(2):[137]–47].
[13] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000.
[15] Fukumizu K. Active learning in multilayer perceptrons. Advances in neural information processing systems. 1996. p. 295–301.
[16] Specht DF. Probabilistic neural networks. Neural Network 1990;3(1):109–18.
[17] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre- diabetes. BMC Med Inf Decis Mak 2010;10(1):16.
[18] Dessai ISF. Intelligent heart disease prediction system using probabilistic neural network. Int J Adv Comp Theory Eng (IJACTE) 2013;2(3):2319–526.
[19] Magnin B, Mesrob L, Kinkingnéhun S, Pélégrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 2009;51(2):73–83.
[20] Cao Y, Hu ZD, Liu XF, Deng AM, Hu CJ. An MLP classifier for prediction of HBV-induced liver cirrhosis using routinely available clinical parameters. Dis Markers 2013;35(6):653–60.