Chronic Disease risk prediction using squirrel search algorithm (SSA) and hybrid KNN

K. Saranyadevi[1], **P. Rathiga**[2]

Abstract

Renal failure is the last stage of chronic kidney diseases (CKD), where the renal functions is partially or totally stop working. Existing prediction model focused on the symptoms for CKD possibility that omits the high risk patients. In this study the Electronic medical record (EMR) is consider to analyze the obvious clinical symptoms to predict the renal failure using efficient feature selection approach. The SSA method is adopted to extract the most significant features from EMR and applies the machines learning algorithms to develop an end to end renal failure prediction model. Various neural networks algorithms such as Artificial Neural Networks (ANN), Multilayer Perceptron (MLP), Elman Network and Radial Basis Function Network (RBFN) are applied as classifier and compared their performance with respect to the evaluation metrics.

KeyWords: Renal failure prediction, machine learning, feature selection, CKD classification, Artificial neural network.

DOI Number: 10.14704/nq.2022.20.8.NQ44699 **NeuroQuantology 2022; 20(8): 6750-6758**

[1] Research scholar, Navarasam Arts &Science College for women, Arachalur, Erode, saranyadevi05061991@gmail.com
[2] Assistant professor, PG and Research Department of Computer Science, Erode arts and Science College, Erode

Introduction

CKD is a common disease that worsening the kidney performance and leads to kidney failure where the prevalence of this kind of disease is increasing annually [1]. The earlier detection of CKD is essential that assist the physicians to improve the treatment. It is the leading risk factor for cardiovascular (CV) risk which has highest mortality rate, particularly in the last stage of CKD [2]. Hence this condition urged to develop the early detection model with highest accuracy.

The recent researches on CKD prediction suggest that the outcomes of these studies assist to prevent CKD and increase awareness of this disease among the patients. The researchers have applied machine learning based classifiers: support vector machine, naïve bayes, and sequential mining optimization [3-4] on CKD prediction. The prior studies mainly focused on classification techniques only few studies carried out with feature selection approaches [5-6]. Feature selection is a significant process in improving the overall performance of the classifier by selecting the most relevant features from dataset.

This study focuses on selecting the essential features of CKD to enhance the performance of detection. The SSA is novel heuristic optimization approach inspired by squirrel food search process. Compared to other optimization technique [7], SSA has significant convergence and has efficient search capability [8] so SSA is suitable for feature selection problem. Hence SSA is utilized in this study to select the significant features and applies the Neural Networks algorithms such as Artificial Neural Networks (ANN), Multilayer Perceptron (MLP), Elman Network and Radial Basis Function Network (RBFN) for prediction.

The contributions of this research are

An end to end renal failure detection model is developed using SSA feature selection and machine learning algorithms.

The proposed feature selection efficiently eliminates the irrelevant features from EMR that improves the classifier accuracy.

The neural networks based classifiers established the high accuracy prediction model with less error rate.

The remaining section of this study is structured as: section II describes the literature survey of existing renal failure prediction model with their achieved result and drawbacks. Section III discusses the methodology of the proposed study with architecture and mathematical explanation. The experimental setup, dataset used and achieved result are explained in section IV. Eventually the present work summary is discussed in section V.

Related work

The CKD prediction approach has been studied by various researchers in past decades. This section discusses the recent studies on renal failure prediction using machine learning algorithms.

Asif Salekin and John Stankovic [9] proposed the CKD prediction method with LASSO regularization feature selection and machine learning classifiers. The 24 features are reduced to 12 attributes and achieved great result with random forest classifier. Lasso regression converts the categorical attributes to perform the reduction process. This study achieved 0.993% accuracy with 0.1084 RMSE.

Rady, El-Houssainy A., and Ayman S. Anwar [10] applied the efficient data mining algorithms to identify the stages of CKD from the patient clinical laboratory data. The GFR value is computed form the observed data and the classification performed with SVM, RBF, PNN and MLP approaches. The PNN achieved 99.72% accuracy compared to others.

Njoud AbdullahAlmansour et al [11] made a comparative analysis on kidney disease prediction using ANN and SVM approaches. The parameters of those two algorithms were optimized by fine tuning them. Several experiments were conducted with these parameters and features and achieved great result with 99.75% and 97.75% accuracy for ANN and SVM respectively. Even though ANN performs slightly better than SVM.

Parab et al [12] introduced Backpropagation ANN and partial least square regression approach to

predict the blood urea and glucose in CKD patients. Then PCA is applied to improve the overall accuracy of the work. BP-ANN model is trained and validated using the PCA values. This method achieved 0.69 and 2.06 mg/dl RMSE values for urea and glucose respectively. This methodology achieved 95.96 % and 98.65 % accuracy for urea and glucose respectively with less available spectral information.

Proposed methodology

The Renal risk prediction model with SSA feature selection and neural network classifier architecture is illustrated in figure1. The proposed system is initialized with scaling the EMR data and selects the most significant attributes using squirrel search algorithm. Then the neural network classifiers predict the risk of renal failure with highest accuracy. The dataset is collected from UCI repository. This model is evaluated with necessary performance metrics and compares with other methods.

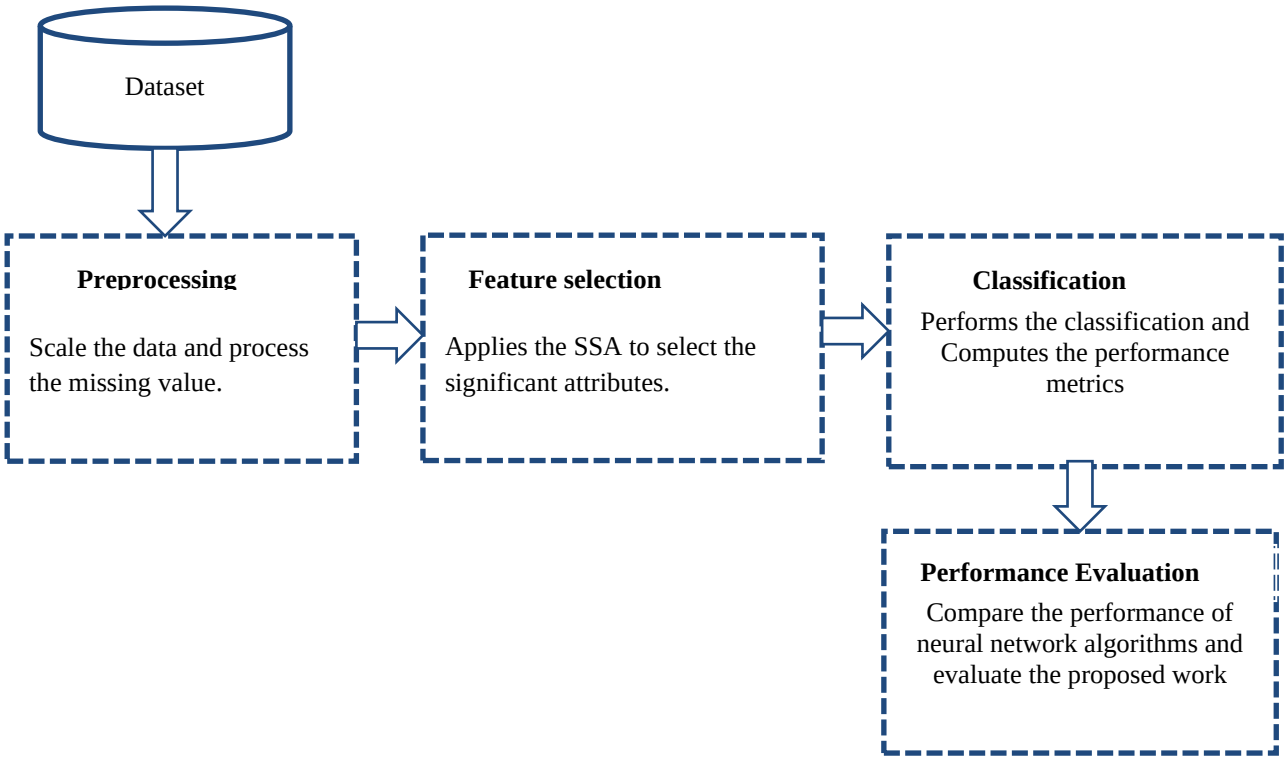

Figure 1. Architecture of the proposed model

Squirrel search algorithm (SSA)

The search process of feature selection based on SSA begins when the squirrel start searching for their food. This process is performed with several area migration. Similar to other natural inspired approach, SSA also forms the migration path based on the fitness values. This strategy will change according to the weather condition in order to increase the likelihood of survival.

Random Initialization

The initial position is assigned using uniform distribution which is described in following equation

$$p_i = p_{min} + U_{(0,1)} \times (p_{max} - p_{min}) \quad (1)$$

p_i denotes the position of all squirrels where p_(i,j) denotes the jth dimension of ith squirrels. U(0,1) represents the uniformly scattered random numbers ranges from 0 to 1. p_max and p_(min)are lower and higher limits of ith squirrel. The objective function of all squirrels were represented by following matrix

$$sof_{i,z} = \begin{bmatrix} sof_{1,1} & sof_{1,2} & \cdots & \cdots & sof_{1,d2} \\ sof_{2,1} & sof_{2,2} & \cdots & \cdots & sof_{2,d2} \\ & \vdots & \vdots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & \vdots \\ sof_{n,1} & sof_{n,2} & \cdots & \cdots & sof_{n,d2} \end{bmatrix}$$

(2)

Where 〖sof〗_(i,z) represents the zth objective function of i th squirrels. The fitness value f=f_1,f_2,⋯,f_z of individual squirrels position is computed by substituting the decision variables into a fitness function

$$f_i = f_i(sdv_{i,1}, sdv_{i,2}, \cdots, sdv_{i,z}) \qquad i = 1,2,\cdots,z$$

Where sdv represents the decision variable, z denotes the population size. Then sorts the quality of food sources in ascending order. The best food source location is identified by minimal fitness value. Then new location can be generated by gliding process which is expressed by following formula

$$sdv_t^{new} = \begin{cases} sdv_t^{old} + d_g G_c (sdv_t^{old} - sdv_t^{old}), & if\ R_1 \geq P_{dp} \\ rl & otherwise \end{cases}$$

Where d_g and G_c denotes random gliding distance [13] and gliding constant, R1 represents a function which returns the value ranges between 0 and 1. The squirrel foraging behavior will change based on the seasonal condition [14]. A seasonal monitoring condition is introduced in SSA

$$S_c^t = \sqrt{\sum_{k=1}^{z} \left(sdv_{t,k}^t - sdv_{t,k}\right)^2}, \qquad t = 1,2,3$$

$$S_{cmin} = \frac{10E-6}{365^{Iter/(Iter_{max})/2.5}}$$

The above mentioned seasonal condition prevents the SSA from being trapped in local optimal solutions. The seasonal condition is checked with S_c^t<S_cmin condition. When the season winter is over squirrel start their searching process by randomly changing their location by

$$sdv_t^{new} = sdv_l + Levy(z) * (sdv_u - sdv_l)$$

The levy distribution is applied in SSA to improve global exploration capability which is expressed in the below formula

$$Levy(x) = 0.01 \times \frac{a_p \times \sigma}{|b_q|^{1/\beta}}$$

a_p and b_q represents the function which gives the values between 0 and 1 , ß. σ is computed by

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\pi\beta/2)}{\Gamma\left((1+\beta)/2\right) \times \beta \times 2^{\left(\frac{(\beta-1)}{2}\right)}} \right)$$

Where $\Gamma(x) = (x-1)$

The algorithm repeated the process by generating the new location with seasonal check and terminates the process when satisfied by maximum iteration. Based on the SSA the attributes in the dataset are filtered and generated the feasible feature set. Then the classification is performed using the neural network algorithm which is discussed in the following section.

Classification using neural network algorithms

Artificial Neural Network

ANN is an efficient method used to solve many applications in medical industry. As in other neural network ANN has three layers: input layers process the input data, output layer provides the relevant output based on the input and the hidden layers connect both input and output layers. The inputs are multiplied by a weight value which inserts a bias and then applies the activation function to generate the required output.

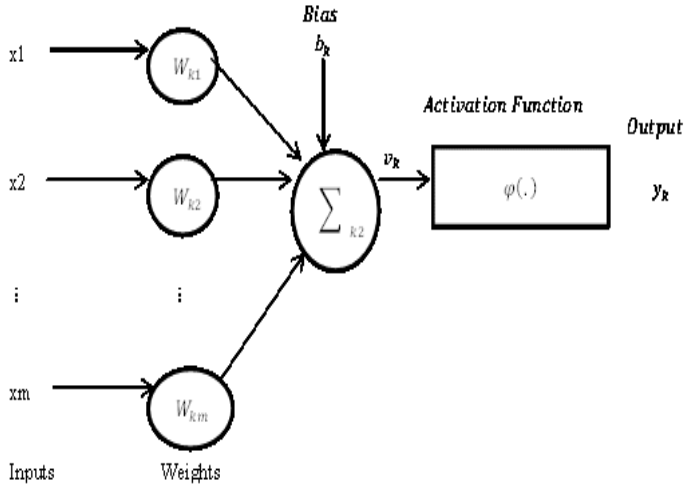

Figure 2. Structure of ANN

Figure 2 illustrate the basic structure of ANN in which the input x_iwere multiplied by weight value w_i among the neuron i and summed by the bias value and passed to hidden layer through a activation function to generate output y_j

$$I_j = \sum_{i=1}^{n} x_i w_{ij} + b_j$$

$$f(I) = {(1 - e^{-2I})}/{(1 + e^{-2I})}$$

$$y_j = f(I_j)$$

Then the out signal is passed to all neuron in hidden layer and generates the output from the input through output layer 〖o^'〗_k

$$o'_k = \sum_{j=1}^{h} y_j w'_{jk} + b'_k$$

Where 〖w^'〗_jk denotes the connected weight among j neurons, 〖b^'〗_k bias value eventually the output is generated with sigmoid activation function.

Multilayer Perceptron (MLP)

Multilayer Perceptron is a feedforward neural network model that maps input onto an appropriate output. It consists of an input layer, one or two hidden layers and an output layer. Nodes that are no target of any connection are called input neurons. Nodes that are no source of any connection are called output neurons. All nodes that are neither input neurons nor output neurons are called hidden neurons. The perceptron in the input layer use linear transfer function and for the perceptrons in the hidden layer and the output layer use sigmoidal functions. The computations performed by each layer are shown below.

The input layer receives the values and serves to distribute the values to the next layer and it does not perform a weighted sum or threshold. It shows that the activity of neurons in the input layer represents the raw information that is fed into the network.

The activity of neurons in the hidden layer is determined by the activities of the neurons in the input layer and the connecting weights between input and hidden units.

The activity of the neuron in the output layer depends on the activity of neurons in the hidden layer and the weight between the hidden and output layers.

Elman Neural Network (ENN)

ENN is a feedback neural network (FNN) approach widely used for prediction in medical data which has four main layers input, hidden, bearing and output layers. The connection of these layers is similar to FNN. The input layer is responsible for transmission where the output layers focus on weighting. The hidden layer has both linear and nonlinear excitation functions by default it set to nonlinear function and used sigmoid. The function of bearing layer is to remember the output of hidden layer which is also called one step delay. The mathematical expression of Elman neural network is

$$O(k) = t\big(w_3 x(k)\big),$$

$$N(k) = f\left(w_1 x_c(k) + w_2\big(u(k-1)\big)\right),$$

$$N_c(k) = N(k-1)$$

Where O(k) represents the output node vector, N(k) represents the nodal element vector, u(k-1) denotes the input vector, x_c (k) is feedback state vector and the connection weight among input to hidden layer is represented by w_3, t() is the transfer method , w_1 weight among connecting layer to hidden layer. f() is the transfer function of hidden layer.

Radial basis Neural Network (RBFN)

The RBFN is one of the feed forward neural networks which have universal approximation capabilities. This network design is viewed as curve fitting approximation problem. The RBFN has 3 layers where input layer is formed with source node. The next node is hidden that applies the nonlinear transformation. The last layer is output layer which is linear focus on supplying the response to activation pattern.

To develop the RBFN, the activation function of invisible layer is set as fixed. Particularly the center of locations is selected randomly from dataset. This network use Gaussian activation which is expressed as ∅_j (x)=e(-‖y_j-ξ_i ‖^2/2σ_j^2) where y_j is the center and σ_j is width, j ranges from j=1,2,⋯,c,. C is the count of center. The algorithm steps to train the network

is as follow

Table 1: Algorithm steps to train the network:

Select the center for the network.

Initialize the center with random data.

Set $E_t = 0$

Select the IO parameters ξ_i^μ, ξ_k^μ $\mu = 1, 2, 3, \ldots n$ and $i = 1, 2, 3, \ldots p$, K-output feature

Calculate hidden layer with Gaussian activation function

Assign the output $O_k = {}^{1}/_{1 + e^{-\Sigma w_{kj} v_j}}$

Measure the error value

Compute the output layer weight as

$$\partial_k = (O_k - \zeta_k) \times O_k \times (O_k - \zeta_k)$$

$$\Delta W k_j = \partial_k \times v_j \times \alpha \times \aleph$$

where α and $\aleph$ represents learning rate and momentum argument

$$W_{kj}^{new} = W_{kj}^{old} + \Delta W_{kj}$$

Check $E_t > E_{min}$ then go to step 4

Dataset

The dataset utilized in this study is available in UCI web source which has 361 patient data with 25 attributes (11 numeric, 14 nominal). The class attribute represents the presence and absence of CKD. The mean of each attribute is considered to remove the missing value present in the dataset. Then the dataset is normalized to form a uniform range by scaling the numeric data with 400 instances for further prediction process.

Result and Discussion

The present model is implemented using python software with keras packages. The kidney patient dataset is collected from UCI repository. The missing values in the dataset are processed by their mean values and the significant features are selected and classified with neural network algorithms. The proposed model is evaluated with confusion matrix, precision, recall, f1 score, accuracy RMSE, Kappa and matthews correlation coefficient value. These metrics are computed using the values of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The definition of above measurement is described as follows,

TP = count of correctly identified as CKD affected person

FP = count of incorrectly identified as CKD affected person

TN = count of correctly identified as not affected by CKD

FN = count of incorrectly identified as not affected by CKD

The performance metrics values are computes with above mentioned measures and listed in below table. Table 1 gives the information of the selected attributes using SSA feature selection algorithm. Table 2 describes the values obtained for each performance metrics which shows the efficiency of the proposed prediction method.

The efficacy of the present study is evaluated with SSA feature selection method and machine learning classifier and tabulated the achieved result.

Table 1: Feature selection result

Dataset	Total Attributes	Feature selection Algorithm	Selected Features	Selected Features Details
UCI CKD dataset	26	squirrel search algorithm (SSA)	17	Age, bp, sg, al, su, rbc, bgr, sc, sod, pot, pcv, wc, rc, htm, dm, pe and ane

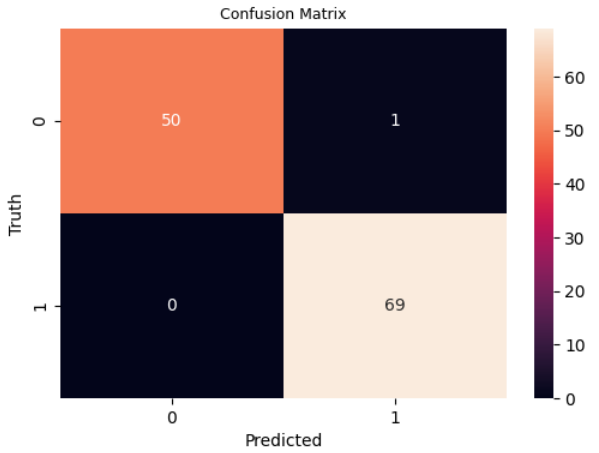

Figure 3 (a). Confusion Matrix of ANN

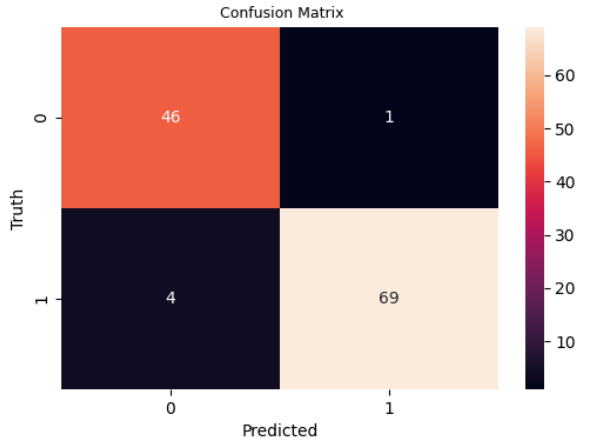

Figure 3 (b). Confusion Matrix of ANN

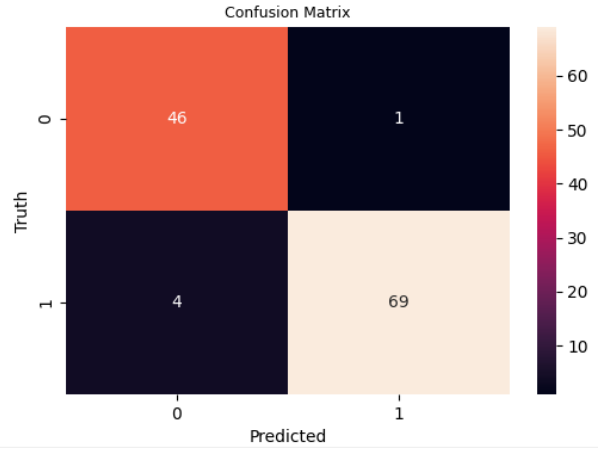

Figure 3 (c). Confusion Matrix of ANN

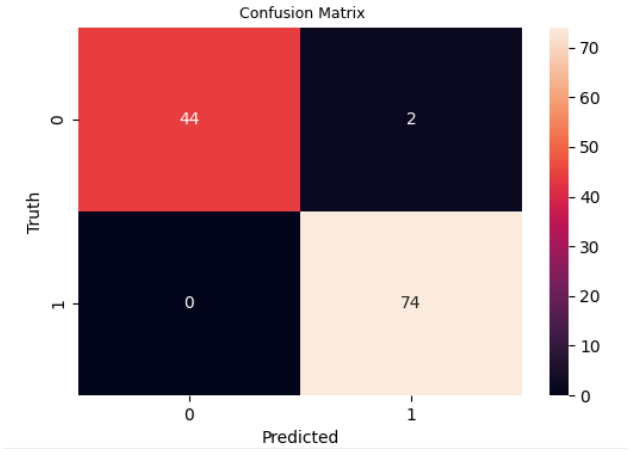

Figure 3 (d). Confusion Matrix of ANN

The Confusion matrix of four machine learning algorithm is illustrated in Figure 3. Four different values such as true positive, true negative, false positive and false negative is present in the confusion matrix. The other performance metrics (precision, Recall and F1 score) can be computed using these values which is given in table 2. The proposed feature selection algorithm is evaluated with four best performing machine learning classifier (ANN, MLP, EN and RBFN). The combination of SSA with four classifier evaluation report is illustrated in following table and figures.

The Evaluation result showed that the combination of SSA with ANN provides the higher accuracy of 99.16% and less error rate of 0.09 compared to other three approaches. The second highest accuracy is achieved by MLP with 98.47% and then the RBFN achieved 97.5% eventually EN got 95.83% accuracy.

Table 2. Performance comparison of Neural network algorithms

	Precision	Recall	F1score	Accuracy	RMSE	matthews_corrcoef	Kappa
ANN	99.18	99.28	99.16	99.16	0.09	0.983	0.982
MLP	98.39	98.57	98.33	98.47	0.12	0.966	0.965

EN	95.91	95.28	95.81	95.83	0.204	0.914	0.913
RBFN	97.58	96.15	97.47	97.5	0.15	0.943	0.941

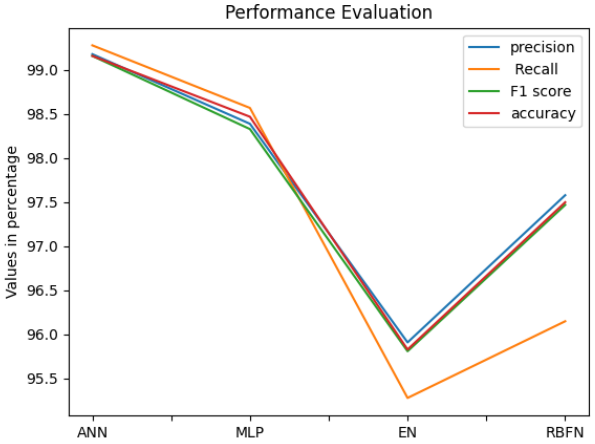

Figure 3. Performance comparison without feature selection

Figure 3 illustrated the performance comparison of machine learning methods by applying the SSA feature selection. The ANN achieved 99.18% precision, 99.28% Recall, 99.16 % F1 score and 99.16% accuracy compared to other methods. MLP achieved second greatest performance (98.39% precision, 98.57% Recall, 98.33% F1 score and 98.47% accuracy) compared to MLP and RBFN. Figure 6 represents the Root mean square error value as mentioned previously the ANN got less error rate of 0.09%.

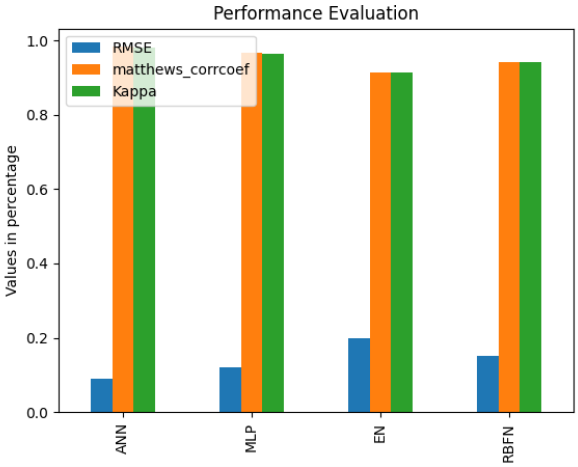

Figure 4. Performance comparison without feature selection

The report in figure 4 illustrate that the proposed model of CKD prediction provides the efficient result with respect to RMSE, Kappa and matthews correlation coefficient. The error value obtained by the neural network algorithm shows that ANN achieved very less error rate in CKD prediction compare to other NN approaches. Likewise other metrics kappa and matthews correlation coefficient also better in ANN. From the evaluation report it is clear that the proposed CKD risk prediction model achieved great result with SSA feature selection approach. Among the applied neural network approach the ANN is perform better with SSA method.

Conclusion

CKD is the severe disease which required the earlier diagnosis system to avoid mortality. In this paper the feature selection and classification is considered for CKD prediction. The SSA is adopted to eliminate the unwanted features and neural network based classifier is executed several times to discover the best feature combination for diagnosis. The SSA with Artificial neural network performs better and provides the efficient result on UCI dataset with less error rate. The ANN achieved 99.18% accuracy with 17 attributes. This combination of attributes provides higher accuracy than the other features set. This study can be further enhanced with hybrid approach to enhance the performance of proposed feature selection approach to reduce irrelevant features in the selected 17 attributes. Also this prediction model can be applied for other diseases like diabetes, heart disease and etc.,

References

1. GBD Chronic Kidney Disease Collaboration, Global, regional, and national burden of chronic kidney disease, 1990-2017: s systematic analysis for the Global Burden of Disease Study 2017, Lancet, 10225 (395) (2020 February 29), pp. 709-733.
2. Sarnak MJ, Levey AS, Schoolwerth AC, et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. Circulation. 2003;108: 2154-216.

3. Bai, Q., Su, C., Tang, W. et al. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep 12, 8377 (2022).
4. D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1305-1310.
5. Pratibha Devishri, S., O. R. Ragin, and G. S. Anisha. "Comparative Study of Classification Algorithms in Chronic Kidney Disease." International Journal of Recent Technology and Engineering (IJRTE) 8, no. 1 (2019): 180-184.
6. Gunarathne, W. H. S. D., K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)." In 2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE), pp. 291-296. IEEE, 2017.
7. Liu, Zhuoran, Fanhao Zhang, Xinyuan Wang, Qidong Zhao, Changsheng Zhang, Tianhua Liu, and Bin Zhang. "A discrete squirrel search optimization based algorithm for Bi-objective TSP." Wireless Networks (2021): 1-15.
8. Wang, Y., & Du, T. (2019). An improved squirrel search algorithm for global function optimization[J]. Algorithms, 12(4), 80.
9. Salekin, Asif, and John Stankovic. "Detection of chronic kidney disease and selecting important predictive attributes." In 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 262-270. IEEE , 2016.
10. Rady, El-Houssainy A., and Ayman S. Anwar. "Prediction of kidney disease stages using data mining algorithms." Informatics in Medicine Unlocked 15 (2019): 100178.
11. Almansour, Njoud Abdullah, Hajra Fahim Syed, Nuha Radwan Khayat, Rawan Kanaan Altheeb, Renad Emad Juri, Jamal Alhiyafi, Saleh Alrashed, and Sunday O. Olatunji. "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study." Computers in biology and medicine 109 (2019): 101-111.
12. Parab, Jivan, Marlon Sequeira, Madhusudan Lanjewar, Caje Pinto, and Gourish Naik. "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Blood Urea and Glucose in CKD Patients." IEEE Journal of Translational Engineering in Health and Medicine 9 (2021): 1-8.
13. M. Jain, V. Singh, and A. Rani, "A novel nature-inspired algorithm for optimization: Squirrel search algorithm," Swarmand Evolutionary Computation, 2018
14. J. Liang, B. Qu, and P. Suganthan, "Problem definitions and evaluation criteria for the CEC 2014 special session and competition on single objective real-parameter numerical optimization," Zhengzhou China and Technical Report, Computational Intelligence Laboratory, Zhengzhou University, Nanyang Technological University, Singapore, 2014.