Chronic Kidney Disease Classification Using Entropy Based Butterfly Optimization and Improved Artificial Neural Network Algorithm: Preprocessing, Feature Selection, Clustering, and Disease Prediction

M. Lincy Jacquline [1], N. Sudha [1]

Abstract

Background: Chronic diseases, which are the leading cause of death globally, account for most medical expenses. Early detection of chronic diseases is crucial for effective preventative medicine. However, the complex nature of these diseases makes accurate early diagnosis challenging. Artificial intelligence (AI) technology has shown promise in assisting clinicians by automating diagnostic processes through predictive models. Methods: This study proposes a classification framework for chronic kidney disease (CKD) using an Enhanced Butterfly Optimization and Improved Artificial Neural Network (EBO-IANN) algorithm. The framework includes preprocessing using the K-means clustering (KMC) algorithm, feature selection via the EBO algorithm, clustering with Weighted Fuzzy C-means (WFCM), and classification using IANN. Results: The proposed EBO-IANN algorithm was evaluated on the CKD dataset from the UCI repository. The framework demonstrated superior performance compared to existing methods, including Bayesian, Neural Networks (NN), Modified K-means Support Vector Machine (MKSVM), and Enhanced Adaptive Neuro-Fuzzy Inference System (EANFIS). The EBO-IANN achieved higher precision, recall, F-measure, and accuracy. Specifically, it showed a precision of 98.9% and 95.9% for the CKD and Hungarian datasets, respectively, with no misclassified features. Conclusion: The EBO-IANN algorithm effectively enhances CKD classification by optimizing feature selection and classification accuracy. This approach can significantly aid in the early detection and treatment of CKD, potentially improving patient outcomes and reducing healthcare

Keywords: Chronic Kidney Disease (CKD), EBO-IANN Algorithm, Preprocessing, Feature Selection, Clustering, Disease Classification

Significance | This study determined advanced AI techniques for accurate CKD diagnosis, enhancing medical decision-making through efficient data analysis methods.

*Correspondence. M. Lincy Jacquline, Department of Computer Science, Bishop Appasamy College of Arts and Science, Coimbatore, IndiaE-mail id-jacquline1990@gmail.com

Editor Surendar Aravindhan, And accepted by the Editorial Board Jun 12, 2024 (received for review Apr 09, 2024)

Introduction

Most medical expenses are related to chronic diseases, which have also been the leading cause of death globally. It is crucial to identify chronic diseases as early as possible to practice preventative medicine. Because of the elusive and complicated pathogeny of chronic disease, it is challenging for clinicians to make an accurate diagnosis in advance. On the basis of their knowledge and experience, physicians typically base the diagnosis of chronic disease on the records of the physical examination (Ekanayake & Herath, 2020). However, if more and more records of physical examinations are collected, physicians will struggle to provide an appropriate diagnosis in the allotted time. AI technology has significantly changed the medical field, and it may assist doctors with diagnosis by automatically generating diagnostic results using

Author Affiliation.
[1] Department of Computer Science, Bishop Appasamy College of Arts and Science, Coimbatore, India

Please cite this article.
M. Lincy Jacquline, N. Sudha, (2024). Chronic Kidney Disease Classification Using Entropy Based Butterfly Optimization and Improved Artificial Neural Network Algorithm: Preprocessing, Feature Selection, Clustering, and Disease Prediction, Journal of Angiotherapy, 8(6), 1-12, 9606

2207-8843/© 2024 ANGIOTHERAPY, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(http.//creativecommons.org/licenses/by-nc-nd/4.0/).
(https./publishing.emanresearch.org).

prediction models. According to the physical symptom is always connected in clinical practice to several chronic conditions.

If kidneys fail to function properly, kidney disease occurs. When kidney function declines, the body begins to accumulate waste materials and extra water, which leads to several issues. Kidney disease and functional impairments lasting for more than three months are referred to as chronic kidney disease (CKD) (Provenzano et al., 2020). It also refers to a decline in glomerular filtration rate (GFR) of 60 ml/min/1.73 m^2 or less for three months or more. The former often refers to a long-term, progressive deterioration in kidney function caused by a consistent loss in renal volume. Because kidney function is lost gradually, CKD is frequently not identified and treated until it is severe (Elhoseny, Shankar, & Uthayakumar, 2019; Kriplani, Patel, & Roy, 2019). Early detection and treatment of kidney disease can decrease the progression of the condition from its current stage to later stages. Knowing the signs of the disease well is crucial for early disease detection.

Researchers are encouraged to continue research in information extraction from clinical datasets by having access to clinical datasets and knowledge-mining approaches. To help clinicians make decisions, several data mining approaches have been employed to mine rules and construct mathematical models. The development of computerized database systems in this information age aids the improvement of medical science decision-making and diagnosis. Data mining methods and tools are used to analyze clinical records in order to create a knowledge-based system that can help physicians make decisions (El-Sappagh & El-Masri, 2014). Data regarding the patients' present health can be found in the clinical dataset (Nahato, Harichandran, & Arputharaj, 2015). The data comprises information about the patient's profile, physical examination, and laboratory test results. Finding hidden important knowledge from clinical datasets is referred to as "mining" and is done to create clinical expert systems.

A significant issue in the quickly developing field of data mining is classification (Amato et al., 2013). Artificial neural networks (ANNs) are highly suited to handle issues related to biomedical engineering because of their broad variety of applications and capacity to learn complicated and nonlinear correlations using noisy or less precise input. According to the investigation, the Multilayer Feed Forward Network with the backpropagation neural network (BPNN) algorithm employing 15 input attributes provides the maximum accuracy. The networks were trained using the BPNN algorithm with momentum and variable learning rates. Different test data were fed into the networks as input in order to evaluate network performance. Different test data were sent to the network as input in order to examine network performance. It was clear from the neural network design that multilayer perceptron neural networks (MLP NNs) needed a smaller architecture about the number of hidden nodes needed for classification than other NNs. As a result, the number of parameters, such as weights and biases, needed to create an MLP NN is significantly less than those of other methods.

The primary goal of this study is the EBO-IANN algorithm-based CKD classification. Despite all of the research and methodswere created, the accuracy of the CKD classifier is not significantly guaranteed. The shortcomings of the current methods include mistake rates and erroneous classification outcomes. The EBO-IANN algorithm is stated to enhance the whole classification efficiency to resolve such issues. The key contributions of this work include preprocessing, feature selection, clustering, and illness classification. The suggested solution employs enhanced algorithms for the supplied CKD dataset to get better results..

Following this introductory section, the subsequent section is a summary of certain aspects of the literature on preprocessing, feature selection, clustering, and disease classification approaches is provided. In Section 3, the suggested technique for the EBO-IANN system is described in depth. Section 4 provides the results of the study and evaluation. Section 5 summarizes the findings.

2.Related work

In , Brisimi et al (2018) worked to develop data-driven strategies to predict hospital admissions caused by two main groups of chronic diseases, heart disease and diabetes. Predictions are based on an individual's electronic health record (EHR), which details their recent medical history. Predictions are defined as binary classification issues where several machine learning techniques have been considered, such as sparse and nuclear support vector machines (SVM), logistic regressions, and random forests. He used two new techniques to strike a balance between predictability and accuracy, so important in the medical environment:

JCC (Clustering and General Classification), finding hidden patient groups and customizing the classifier for each group, and K-LRT, a technique based on probability ratio tests. For the latter approach, create theoretical out-of-sample assurances. The Boston Medical Centre, New England's largest safety-net hospital system, provided significant data sets for its algorithm validation.

In Yildirim et al (2017) created a neural network classifier for healthcare decision-making regarding chronic renal disease, it investigates the impact of class imbalance in training data. Many applications, such as data mining and decision systems, regularly use neural networks. Popular neural network architectures that can be trained to recognize various patterns include back propagation networks, considering their significances and comparative analyses of sampling techniques for diagnosis of chronic renal failures were carried out utilising multilayer perceptrons using different learning rates.The study proved that learning rates are essential parameters

impacting performances of multilayer perceptrons and sampling techniques enhance accuracies of classifications.

In Aqlan et al (2017) examined the use of analytics and data mining approaches to predict CKD. Many missing data imputation techniques were used since the data contains missing values. The C&RT algorithm was chosen because it minimizes the variability in the imputed data. In order to forecast CKD, six predictive analytics techniques were used, and it was discovered that Random Trees is the most effective technique. Only the classifications ckd and notckd are included in the data utilized in this investigation for the output variable. But the National Kidney Foundation indicates that the ckd class can be further divided into five classes. All Class 0 (notckd) cases were grouped into one cluster (cluster-1) after the data were clustered using K-means into six clusters. Cases of class 1 were divided into 5 clusters with varying percentages. Future research will concentrate on studying and predicting the five stages of CKD using regression and clustering algorithms.

In, Ghosh et al (2020) described the late diagnosis of CKD was a significant problem and many men and women suffer from early screening of the illness though their early identifications can save lives.A trustworthy dataset can also help a machine learning algorithm's analysis procedure identify this dangerous disease's stage much more quickly. The research's entire implementation relied on four reliable approaches:AdaBoost, SVM (Support Vector Machine), LDA (Linear Discriminant Analysis), and GB (Gradient Boosting). These techniques are used using a dataset from the UCI machine learning repository that is available online. GB Classifiers produce results with a predictably high accuracy of roughly 99.80%. Following that, many metrics for assessment of performance were also displayed to validate the right results. The best and most efficient approaches for the work can be selected according to benchmarks.

In Lambert et al (2020) predictedCKDemploying attributes that are nominal and produce outcomes that are comparable to numerical attributes. Correlation-based feature selection (CFS) method is utilized in the classification and prediction to extract key characteristics and categorize them as belonging to or not belonging to CKD. CFS (correlation-based feature selections) were applied on nominal, numeric, and associative data attributes for selecting features. The outcomes of CFS using Incremental, Attained, and Reduced F ranking algorithms on features from original CKD datasets and SMO categorizations of renal disease states were successful. Hence, CFS-SMO was recognized as a tool that could reliably diagnose renal illnesses and help medical professionals reach right conclusions.

In Vashisth et al (2020) used a fully integrated Deep Neural Network is used by a multi-layer perceptron classifier. If the patient has chronic renal disease, to rule it out. The algorithm employed investigates numerous symptoms, such as patients' ages', blood pressures, red blood cell counts, etc., to assist systems categorise accurately. It is trained using a dataset of roughly 400 cases. The findings show that the model can classify the data more accurately than SVM and Nave Bayes Classifier, with a test accuracy of 92.5.

2.Methods and Material

Early-stage CKD are predicted using EBO and IANN algorithms in this work The primary contributions of this work are preprocessing, feature selection, clustering, and CKD prediction. Fig 2 shows the general framework of the suggestedEBO-IANN algorithm

2.1 Pre-processing using KMC algorithm

CKD medical dataset was pre-processed by KMC to increase accuracy where KMC effectively clustered comparable data based on initial focus of clusters (Ismail et al, 2014). The centroid of the cluster is located using the concept of Euclidean distance. The processes start with random partitions and iterating current cluster centre locations (cluster mean vectors in data), and then (ii) reallocates data points to clusters whose centres are closest to them and halt executions when there are no more clusters. These result in minimizing local inter-cluster variances—sum of squares of changes between data features and corresponding cluster centers. A KMC algorithm example is shown in Fig 3.

Its two key benefits are the linear execution time and the effectiveness of the K-means implementation. As there are more classes, there are an equal number of clusters. The formula below is used to obtain clusters' centroids:

$$d(i,j) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Where x_i and y_i are n-spaced Euclidean points

2.2 Algorithm 1: KMC algorithm

To begin the clustering process, first select a subset of data points from the CKD dataset. Next, specify the initial cluster centers, denoted as μ_1, ..., μ_k, by determining and positioning these centers at k distinct data points. Once the cluster centers are set, assign data points to clusters at random. Using the specified formula, calculate the distance between each data point and the nearest cluster center to determine the closest cluster. Assign each data point to its nearest cluster and update the cluster centers by recalculating the cluster averages. Throughout the process, identify and eliminate any missing or incorrect values. The process continues iteratively until no new assignments occur, at which point the clustering procedure concludes.

Versions of the original dataset that lack certain features are eliminated. Dataset was split into a version with complete samples and and other version with partially complete samples and missing values. KMC is used to the whole instance collection to produce full instance assembly. As a result, each instance is examined independently, and any attributes that are lacking are filled up with plausible values. On implementing KMC clusters were generated

from the dataset values and newly added instances were examined to determine their classifications into appropriate classes and the procedure repeated. The following value are given and compared until proper cluster instances in incorrect clusters are located. Hence, ths usage of KMC algorithm in preprocessing procedures effectively enhance CKD classification accuracies.

2.3 Feature selection via EBO) algorithm

EBO, an updated algorithm that draws inspiration from nature and imitates the food-seeking (more accurate with certain attributes) and mating behaviors of butterflies, selects best features from CKD medical dataset and solves disease diagnostic classifications. EBO is primarily inspired by how butterflies utilize their sense of smell to select the most advantageous traits to use to locate their nectar partners [15]. According to research, it has been discovered that butterflies can identify the source of smell with a high degree of classification accuracy.

A butterfly's fitness (precision of categorization) changes when it goes from one location to another since it is correlated with the strength of the fragrance it emits. The three key concepts of detection, processing, and technique are the cornerstones of the EBO algorithm. For the best feature selection, consider sensory modalities (c), stimulus intensities (I), and power exponents (a) . I has to do with how well (accurately) the EBO algorithm chose the features it did from the set of medical data. The EBO algorithm creates fragrances based on the strength of the physical stimulation using these concepts and equation (9).,

$$f = cI^a \qquad (2)$$

here f implies perceived magnitudes of fragrances, i.e., the sensory modality produced by classifier accuracy is the potency with which other butterflies can sense the perfume. In stands for stimulus intensities, and an implies mode-dependent power exponents. A and C are so located between [0, 1]. If a = 0, on the other hand, it means that other butterflies are unable to smell them. As a result, the parameter a controls how the algorithm behaves.C is a vital parameter for determining the EBO algorithm's speed of convergence. It is another significant parameter. To explain the themes in relation to a search method, the following characteristics of butterflies have been modified:

According to theory, all butterflies should release a scent that attracts other butterflies with similar characteristics.

Seconly, Every butterfly will migrate at arbitrary or the direction of the one which has best smell.

Thirdly, A butterfly's stimulus intensity is influenced or determined by the landscape in some way.

Initialization, iteration, and final stage are the three steps of an EBO. Every time the EBO is executed, the initialization phase is carried out first, followed by the iterative feature search, and the algorithm is terminated after the best option for the best choice has been identified. The initialization stage of the EBO method involves calculating the solution space and classification precision. Values for EBO-related parameters are also assigned. The placements of butterflies (features), along with their fitness and scent values, are created at arbitrary in the feature selection search space. The algorithm begins the iteration phase after completing the initialization step. The butterflies in the feature selection solution space are all moved to novel places throughout iterations, and their accuracies of classification are considered. In procedures, all butterflys' fitnesses are first computed at various locations in solution spaces. Next, utilizing equation (3), These butterflies will emit scent wherever they are present. The butterfly moves toward the fittest solution (g*)(optimal features) in global searches illustrated by equation (3),

$$x_i^{t+1} = x_i^t + (r^2 \times g^* - x_i^t) \times f_i * ECE_W \qquad (3)$$

here x_i^t are solution vectors x_i for ith butterflies in iterations t. g^* implies present best feature solutions in iterations amongst available solutions. Fragrances of ith butterflies are indicated by f_i and $r \in [0, 1]$ are random numbers. Local searches are depicted as equation (4),

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i * ECE_W \qquad (4)$$

where x_j^t and x_k^t are jth and kth butterflies in feature selection solution spaces. If x_j^t and x_k^t belong to same swarms and $r \in [0, 1]$ implies random numbers, then equation (11) results in local random walks. To get the best attribute selection from the dataset, butterflies can forage and mate both domestically and abroad. EBO shifts from general global search to intense local search with variable probability p. The iteration phase continues up until the halting condition is not satisfied. This method produces the ideal solution in the best possible physical condition at the conclusion of the iteration. The EBO technique selects the ideal number of features in the CKD medical dataset using additional feature weights in equation (4). The EBO approach concentrated on employing the best possible feature selection across the provided CKD medical dataset to increase the classifier's accuracy. By minimizing the distance among two sample distributions, an optimization issue is solved, and the best probability distribution parameters are then obtained. This technique is known as cross entropy (CE). The CE approach provides high robustness, outstanding adaptability, and good global search capabilities.

$$CE = \frac{1}{N}\sum_{i=1}^{N} I_{s<r} \frac{f(x^i, v)}{g(x^i)} \qquad (5)$$

where x^i denotes random samples from $f(x; v)$ with importance sampling densities $g(x)$. Kullback-Leibler divergence or cross entropy is used to quantify the separation between two sample distributions in order to estimate the optimal meaningful sampling density.

The algorithm 2 illustrates the total steps included in the suggested EBO algorithm. Initial populations for algorithm 2 are produced

using feature counts in medical datasets (Step 1), and I_i (stimulus intensity) at x_i (Step 2) are computed based on sensor modalities c, power exponents a (Step 3). The classification accuracy is the source of these variables. The process then begins with halting criteria (Step 4), after which each butterfly's scent value is calculated (Step 6). Finding the population's best feature came next (Step 8), and then r was developed (Step 10). If $r < p$then use equation (4) to go in the direction of the best butterfly, or equation (12) to move arbitrarily. Then, in steps 17 and 18, you update a value and evaluate people in light of their new standing. Lastly, use the end while command (Step 19) to finish the procedure. The flowchart of the suggested Entropy Butterfly Optimization (EBO) system is illustrated in Fig 4.

Algorithm 2: EBO algorithm

Input: CKD medical dataset

Objective functions: Classifier accuracy, $f(x), x = (x_1, x_2, \dots, x_{dim})$ $dim = no.of\ dimesnions$

Output: Selection of optimal features (age, blood pressure, sugar level, hemoglobin and so on)

Set up initial populations of n butterflies $x_i = (i = 1,2, \dots, n)$ using dataset's features

Stimulus Intensities I_i at x_i are obtained in classification accuracies $f(x_i)$

Describe sensor modalities c, power exponents a and switch probabilities p

Do until stopping criteria not met

1. For butterflies f in population do
2. Compute fragrances f with equation (3) and generate weights using equation (5) based entropies
3. End for
4. Find best butterflies
5. For butterflies f in population do
6. Generate random numbers r
7. If $r < p$ then
8. Move towards best butterflies (optimal features) using equation (3) and generate weights via entropies of equation (5)
9. Else
10. Move arbitrarily using the equation (4)
11. End if
12. End for
13. Update the value of a
14. Estimate individuals(features) relying on their new position
15. End while
16. Output best obtained solutions

Clustering using WFCM

To decrease misclassification, class labels may need to be predicted after feature selection. To forecast class labels, this study clusters data using WFCM. Each data point in a fuzzy clustering analysis may be associated to more than one cluster. In cluster analysis or clustering, data points are assigned to clusters where items within clusters are similar and items in separate clusters are dissimilar. Clusters are found using similarity measurements that account for distance, connection, and intensity. Depending on the data or the application, several similarity metrics are used [17] [18].

Disadvantages of FCM

Some characteristics in high-dimensional signals may be irrelevant or relevant, although they could have different clustering significance. These features must be incorporated into the clustering process for better clustering. To address these problems, a weighed FCM clustering approach is proposed.

WFCM Clustering

When objective functions J achieve minimal values, Fuzzy c-means (FCM) assists in distributions of data X = {x1,…, xi,….xn} (1≤i≤n) into c clusters based on membership degree matrices U=(uti)c*n. FCM split the data into clusters based on membership degree matrices holding orders of members when objective functions J achieve lowest values . Tablets U=(uti)c*n. Xi of X have dimensions p, and uti denotes the degree to which the sample xi is a member of the cluster centre vt. Clusters c are identified by cluster centres V={v1,….vt,….vc} 1≤t ≤c where V is selected randomly in the beginning phases. The adhesion level is then determined as follows:

$$u_{ij} = \frac{1}{\sum_{z=1}^{c}(d_{ti}/d_{zi})^{2/(m-1)}} \quad (6)$$

here dtirepresents Euclidean distances of samples xi from cluster centers vt and m denotes power exponents. Cluster centres are computed using:

$$v_t = \frac{\sum_{i=1}^{n} u_{ti}^{m} d_{ti}}{\sum_{i=1}^{n} u_{ti}^{m}} \quad (7)$$

The objective function J is stated below

$$J = \sum_{i=1}^{n} \sum_{t=1}^{c} u_{ti}^{m} d_{ti}^{sy\,2} \quad (8)$$

The weighted Euclidean distance serves as the foundation for feature-weight learning. The widely utilized Euclidean distance is dij, while the weighted Euclidean distance is d_{ij}^{w}, as seen below:

$$d_{ij}^{w} = \sqrt{\sum_{k=1}^{s} w_k \left(x_{jk} - v_{tk}\right)^2} \quad (9)$$

The following is an overview of the resulting FFCN algorithm.

Step 1: Set the threshold value and the maximum counts of clusters c and ensure use of m as the proper constant.

Step 2: List the FCM's members and hubs.

step (3). Calculate (ij) in accordance with (6) Let's thus update vt based on (7) using the new computation u_(ij).

Step 4: Applying (8), estimate the goal function J.

Step5:If the goal function J's difference between two adjacent computed values is smaller than the stated threshold, the process converges or terminates. If not, move on to step 3.

Based on the above procedure normal subjects and Kidney Disease subjects are separately by different cluster

CKD Classification using IANN

IANN technique is used to classify CKD. ANN is employed for knowledge acquisition through learning. Input, hidden, and output layer are the three stages of an ANN. The input layer gathers input data features, and after processing them, produces 'n' inputs. These procedures follow a set of weights. Weights are the data that neural network issues are solved with . After some effective hidden extraction, the hidden information is taken from the input layer and sent to the output layer. The classification of the CKD medical dataset in this case uses IANN. IANN was used to train the balanced dataset, and state features were tested to categorize the features. The ANN is improved with Multilayer Perceptron (MLP) via sigmoid function which is called as IANN. Fig 5 shows the ANN architecture

Input Layer - The network's input layer carries the chosen characteristics of the CKD data. Initially, the data is somewhat undeveloped.

Hidden Layer – The fundamental function of the hidden layer is to transform the input layer's raw dataset data into something that the output layer can use. One or more hidden levels may be providein the IANN architecture.

Output Layer: The output layer processes data given from hidden layers for achieving desired outcomes (better classifier accuracies and shorter execution times).

MLP Feedforward Neural Network (FNN) architecture, in which neurons are arranged cascade-wise, is the most often used FNN model. At least two layers make up MLP. The input of the layer i+1 neuron is the output of the ith layer in MLP; there is no connection between the neurons that comprise a layer. The nodes in the input layer is inversely proportional to counts of features in the input vector, and node counts in output layers are inversely proportional to output layer counts.

$$Y_n = f(\sum_{m=1}^{h}(w_{nm}, f(\sum_{l=1}^{i} v_{ml}X_l + \theta_{vm}) + \theta_{wm} \quad (10)$$

$$n = 1, \ldots\ldots, o$$

where Y_n represents outputs of nth nodes in output layers, X_l signifies inputs of lth nodes in input layers, Nodes m in hidden layers and n in output layers are connected by connective weights wnm. v_{ml} stands for connective weights amongst nodes l in input layers and m in hidden layers and θ_{vm} and θ_{wm}denote bias terms or transfer function threshold f of nodes m in hidden layers and n in output layers

In IANN, if the weighted sum of the inputs is greater than a programmable cutoff value, also known as an activation function, the perceptron model transmits the output 1. Each neuron's output is the weighted sum of its inputs, with its bias. 'w' and 'x' are the weights and input neuron correspondingly

$$\sum_{i=1}^{m} bias + (w^i x^i) \quad (11)$$

One of the functions included is the Sigmoid function, which is utilized by the activation function.

$$f(x) = sigmoid = \frac{1}{1+\exp(-x)} \quad (12)$$

Connection weights and offsets of neurons are included in network weighnts. The most effective technique to acquire the intended outcomes from the y entry is through neural network training, which are processes of updating network's weights and figuring out the right values for the weights and biases. The CKD economics is offered.

Algorithm 3: IANN

Input: Selected features(age, blood pressure, sugar level, hemoglobin and so on)

Output: Better classification results for CKD medical dataset

1. Procedure IANN (inputs, neuronss, repetetions)
2. Generate input database
3. Input←databases with all possible combinations
4. Train IANN
5. For inputs = 1 to end of inputs do
6. For neurons =1 to n do
7. For repetetions = 1 to n do
8. Train IANN
9. IANN-storages ←save highest accuracy feature values
10. End for
11. End for
12. IANN-storage←save best predictions of IANN based on inputs
13. End for
14. Return IANN-storage → Result with best classification of IANN for every feature combinations

Result

The experimentation done on the suggested model is examined in this section. MATLAB is utilized to carry out the model's implementation. Comparison of the existing NN, MKSVM, EANFIS algorithms and the proposed EBO-IANN algorithm are done according to precision, specificity, accuracy and F-measure for the CKD Data Set from UCI repository (https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease). It contains 25 Attributes, 400 number of Instances.

Precision

It is calculated as,

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}} \quad (13)$$

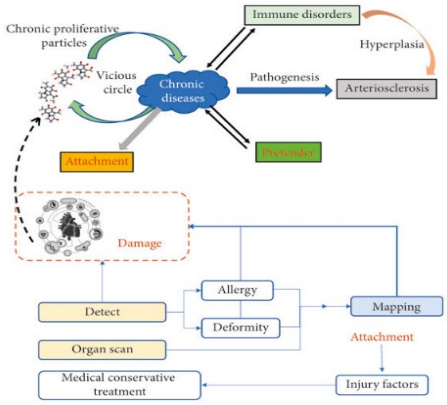

Figure 1 Framework of CKD classification

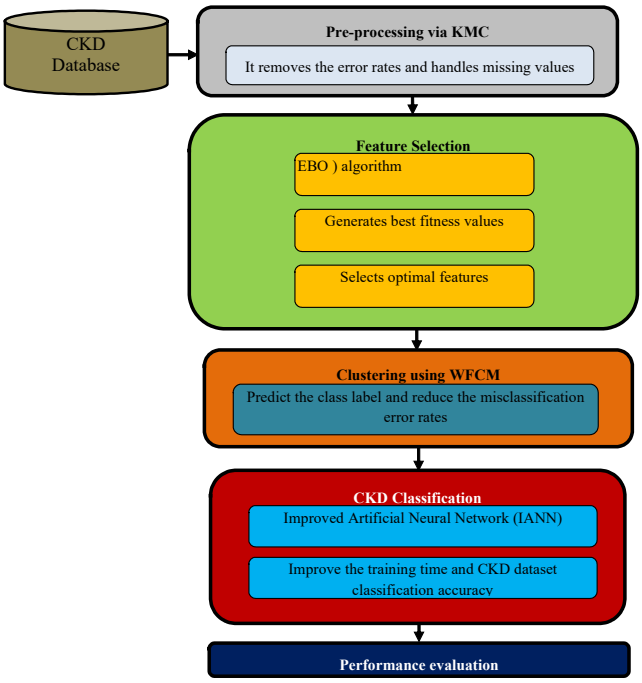

Figure 2. Overall block diagram of the proposed system

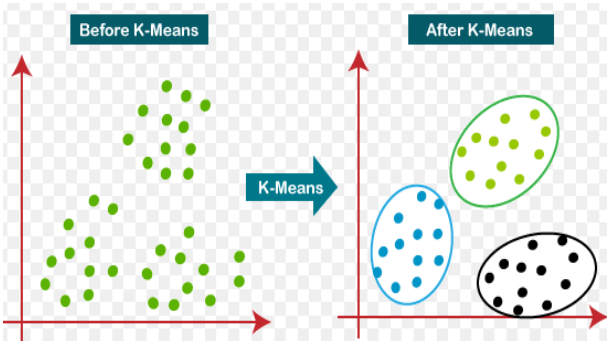

Figure 3. Example of KMC algorithm

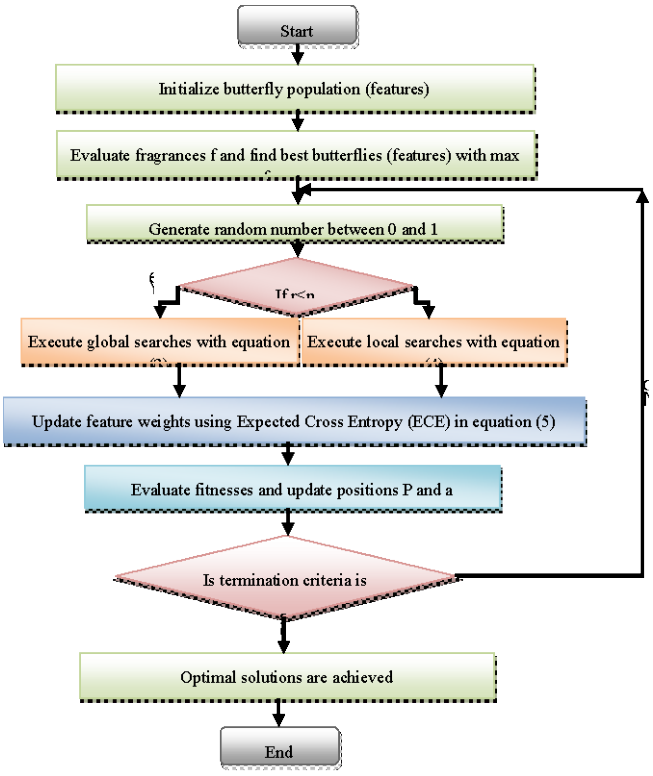

Figure 4 flowchart of EBO algorithm

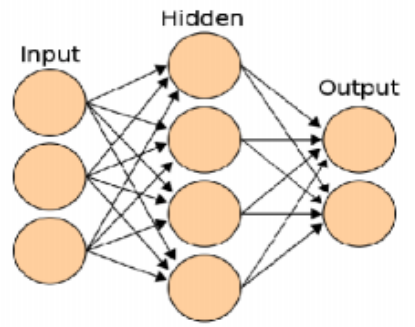

Figure 5 Architecture of ANN

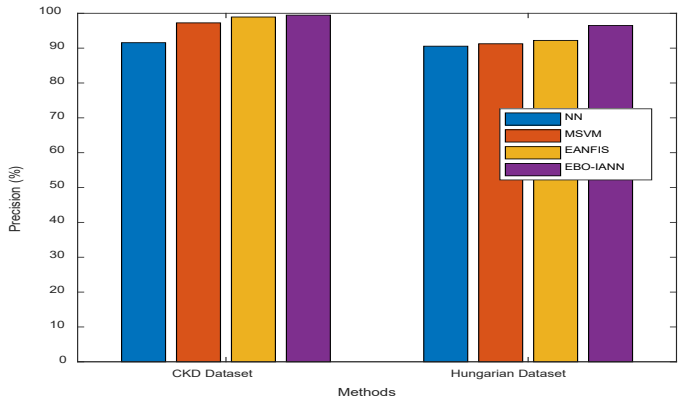

Figure 6 Precision

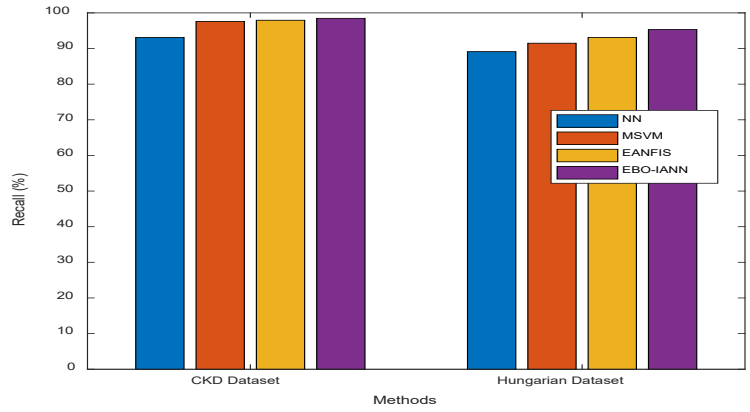

Figure 7 Recall

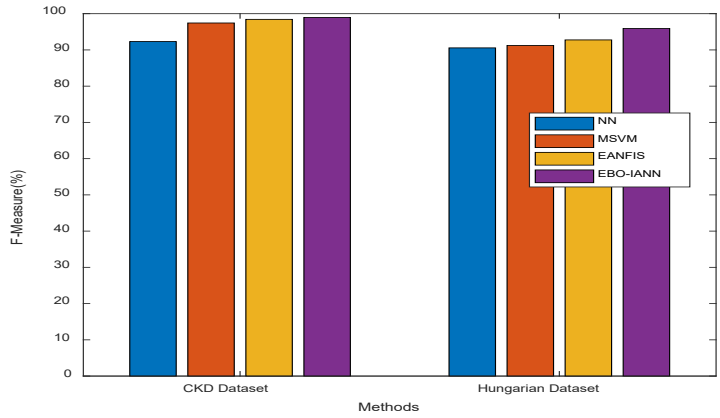

Figure 8 F-measure

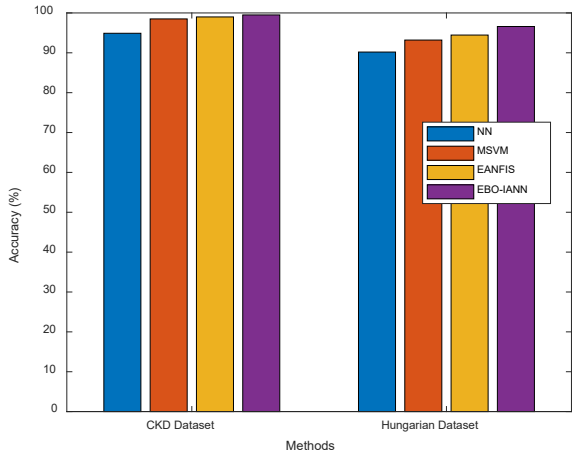

Figure 9 Accuracy

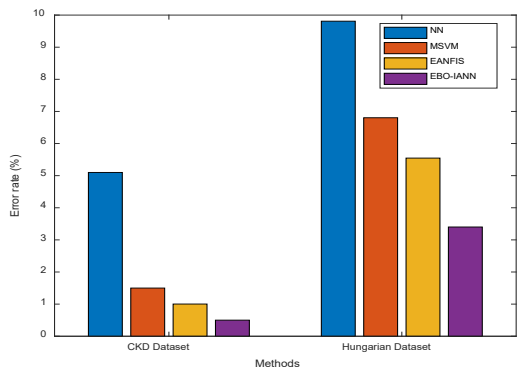

Figure 10 Error rate

Recall is an evaluation of correctness or quality even if it is an estimate of thoroughness or quantity. Algorithms with high accuracy are frequently more beneficial than those with low precision. Divide false positive counts by total true positive counts, for instance, to calculate accuracy.

The assessed accuracies of suggested and existing methods are depicted in Figure 6 where x-axis represents methodologies, while precision values are pl;otted in the y-axis. Conventional techniques include Bayesian and NN algorithms, MKSVM, and naïve EANFIS provide lesser accuracy for specific CKD and Hungarian medical data sets, whereas the new EBO-IANN methodology delivers higher accuracy. The suggested method improves precisions by selecting more relevant information. The findings lead to the conclusion that the proposed EBO-IANN method enhances CKD classification performance by using the best features.

Recall

The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{T_p}{T_p+F_n} \quad (19)$$

The pertinent document counts found by searches divided by total relevant documents counts in existence is called recall, while relevant document counts found by searches divided by total documents found counts in searches are known as precision.

The specificity values in comparisons were assessed using both existing and proposed methods, as seen in Figure 7 above. Specificity data are displayed on the y-axis together with the methods utilised for the x-axis. For the given CKD and Hungarian medical datasets, the suggested EBO-IANN algorithm offers more recall, whereas existing methods like NN, MKSVM, and EANFIS algorithms offer less recall. The training stability is enhanced by EBO-IANN. The findings show that the suggested EBO-IANN method improves CKD and Hungarian dataset classification accuracy by using the best features.

F-measure

F1-Score is described as:

$$F1-\text{score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \quad (15)$$

The comparative values for the F-measure metric utilizing the existing and suggested algorithms are assessed from Fig. 8. For the provided CKD and Hungarian medical datasets, the suggested EBO-IANN algorithm yields higher F-measure compared to the current NN, MKSVM, and EANFIS approaches. The proposed classifier demonstrated that F1 score of 98.9 and 95.9% for CKD and Hungarian medical datasets respectively, without any incorrectly identified features. Also, the findings depict that the existing NN, MSVM, EANFIS techniques produce lower f-measure for both datasets. EBO provides optimal features and on CKD and Hungarian medical datasets, the suggested approach offers improved classification accuracy and greater efficiency.

Accuracy

It is determined by dividing the total number of actual classification parameters $T_p + T_n$) by the total number of classification parameters ($T_p + T_n + F_p + F_n$). Its definition is the total precision of the approach.

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \quad (16)$$

here Tp is true positive, Fp is false positive Tn is true negative and Fn is false negative

The accuracy of the comparison measurement was assessed utilising existing and proposed techniques, as seen in Figure 9. The precise value is shown on the y-axis, while the data set and procedure are utilised for the x-axis. The suggested EBO-IANN algorithm outperforms existing methods, such as NN, MSVM, and EANFIS approaches, for a particular CKD and Hungarian medical datasets. Pre-processing techniques are used to remove noise and fill in missing values to enhance classification accuracy. The outcomes show that the existing NN, MSVM, and ENFIS approaches provide lower accuracy results and proposed EBO-IANN algorithm provides higher accuracy. As an outcome of the optimal feature selection, the suggested EBO-IANN technique boosts the CKD dataset accuracy.

Error rate

The proposed system is better when it provides lower error rate

Fig 10 compares the error rates of the proposed EBO-IANN approach's and traditional approaches like NN, MSVM and EANFIS. X-axis represents different methods and Y-axis represents the error values. The findings depict that the EBO-IANN algorithm delivers lower error rate of 0.5% and 3.4% for CKD and Hungarian medical datasets respectively, while existing NN and MSVM techniques produce higher error rates

Conclusion

EBO-IANN method is proposed in this paper for improving classifications of CKD medical dataset samples. The four primary components of this study are feature selection, clustering, classification, and pre-processing. The KMC algorithm is used to eliminate noise and fill in missing variables in order to maximise classification performance. Then, EBO is used for feature selection, allowing the selection of pertinent and practical features. Finally, classification is done by using IANN algorithm which provides more accurate CKD dataset classification performance. The class label of tuples is predicted using a set of weights that are learned iteratively. The findings led to the conclusion that, compared to the current methods, the suggested EBO-IANN algorithm offers higher accuracy, specificity, recall, and f-measure. In future work, feature

extraction and hybrid classification algorithm can be developed for the given dataset

Author contributions

M. Lincy Jacquline contributed to the methodology, data collection, analysis, and the initial drafting of the manuscript. Additionally, she was responsible for the conceptualization, supervision, project administration, and review and editing of the final draft. Dr. N. Sudha played a significant role in the methodology and data interpretation, as well as in the review and editing of the manuscript. Dr. Sudha also contributed to data collection and analysis.

Acknowledgment

Author was grateful to their department.

Competing financial interests

The authors have no conflict of interest.

References

Abdulrahman, S., Ali, S., & Khan, N. (2021). Machine learning approaches for the prediction of chronic kidney disease progression. Healthcare Informatics Research, 27(1), 54-62. https://doi.org/10.4258/hir.2021.27.1.54.

Ahmad, F., Nazir, S., & Khan, T. (2019). A novel hybrid method for chronic kidney disease prediction. Neural Computing and Applications, 31, 4931-4944. https://doi.org/10.1007/s00521-018-03937-w.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine, 11(2), 47-58. https://doi.org/10.2478/v10136-012-0031-x.

Aqlan, F., Markle, R., & Shamsan, A. (2017). Data mining for chronic kidney disease prediction. In IIE Annual Conference Proceedings, pp. 1789-1794.

Arora, S., & Singh, S. (2019). Butterfly optimization algorithm: A novel approach for global optimization. Soft Computing, 23, 715-734. https://doi.org/10.1007/s00500-018-3102-4.

Brisimi, T. S., Xu, T., Wang, T., Dai, W., Adams, W. G., & Paschalidis, I. C. (2018). Predicting chronic disease hospitalizations from electronic health records: An interpretable classification approach. Proceedings of the IEEE, 106(4), 690-707. https://doi.org/10.1109/JPROC.2017.2789319.

Chen, J., Zhou, X., & Ma, J. (2020). Deep learning-based prediction of chronic kidney disease using multi-layer perceptron. IEEE Access, 8, 53960-53970. https://doi.org/10.1109/ACCESS.2020.2976767.

Dwivedi, A. K. (2018). Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 29, 1545-1554. https://doi.org/10.1007/s00521-016-2701-1.

Ekanayake, I. U., & Herath, D. (2020). Chronic kidney disease prediction using machine learning methods. In Moratuwa Engineering Research Conference (MERCon) (pp. 260-265). https://doi.org/10.1109/MERCon50084.2020.9185249.

Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific Reports, 9(1), Article 13366. https://doi.org/10.1038/s41598-019-46074-2.

El-Sappagh, S. H., & El-Masri, S. (2014). A distributed clinical decision support system architecture. Journal of King Saud University-Computer and Information Sciences, 26(1), 69-78. https://doi.org/10.1016/j.jksuci.2013.03.005.

Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020). Optimization of prediction method of chronic kidney disease using machine learning algorithm. In 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (pp. 1-6). https://doi.org/10.1109/iSAI-NLP51646.2020.9376787.

Gupta, R., Yadav, A., & Verma, S. (2018). Enhanced chronic kidney disease prediction using ensemble methods. Journal of King Saud University-Computer and Information Sciences, 32(10), 1201-1209. https://doi.org/10.1016/j.jksuci.2018.09.006.

Ismail Bin, M., & Dauda, U. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6, 17.

Jain, R., Kumar, A., & Singh, R. (2019). Predicting chronic kidney disease using machine learning algorithms. International Journal of Advanced Computer Science and Applications, 10(8), 296-303. https://doi.org/10.14569/IJACSA.2019.0100841.

Kriplani, H., Patel, B., & Roy, S. (2019). Prediction of chronic kidney diseases using deep artificial neural network technique. In M. A. Lebbah, A. Benyettou, & M. Sadok (Eds.), Computer Aided Intervention and Diagnostics in Clinical and Medical Images (pp. 179-187). Springer. https://doi.org/10.1007/978-3-030-04061-1_18.

Lambert, J. R., Arulanthu, P., & Perumal, E. (2020). Identification of nominal attributes for intelligent classification of chronic kidney disease using optimization algorithm. In International Conference on Communication and Signal Processing (ICCSP) (pp. 0119-0125). https://doi.org/10.1109/ICCSP48568.2020.9182206.

Liu, X., Xu, Y., & Zhang, H. (2020). A comprehensive review on deep learning techniques for chronic kidney disease prediction. Journal of Healthcare Engineering, 2020, Article 7868123. https://doi.org/10.1155/2020/7868123.

Nahato, K. B., Harichandran, K. N., & Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and Mathematical Methods in Medicine, 2015, Article 460189. http://dx.doi.org/10.1155/2015/460189.

Nair, A. K., Rao, R., & Menon, V. (2018). Hybrid machine learning models for effective chronic kidney disease prediction. International Journal of Medical Informatics, 115, 37-45. https://doi.org/10.1016/j.ijmedinf.2018.04.008.

Nayak, J., Naik, B., & Behera, H. (2015). Fuzzy C-means (FCM) clustering algorithm: A decade review from 2000 to 2014. In Proceedings of the International Conference on Computational Intelligence in Data Mining (CIDM) (Vol. 2, pp. 133-149). https://doi.org/10.1007/978-81-322-2208-8_14.

Patil, P., & Bhise, A. (2019). A comparative study of machine learning algorithms for chronic kidney disease prediction. International Journal of Innovative Research in Computer and Communication Engineering, 7(5), 2340-2348. https://doi.org/10.15680/IJIRCCE.2019.0705085.

Provenzano, M., Andreucci, M., De Nicola, L., Garofalo, C., Battaglia, Y., Borrelli, S., Gagliardi, I., Faga, T., Michael, A., Mastroroberto, P., & Serraino, G. F. (2020). The role of prognostic and predictive biomarkers for assessing cardiovascular risk in chronic kidney disease patients. International Journal of Nephrology, 2020, Article 2314128. https://doi.org/10.1155/2020/2314128.

Qin, J., Fu, W., Gao, H., & Zheng, W. X. (2016). Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory. IEEE Transactions on Cybernetics, 47(3), 772-783. https://doi.org/10.1109/TCYB.2016.2526683.

Sharma, N., Mittal, A., & Garg, S. (2017). An effective approach for chronic kidney disease prediction using machine learning techniques. International Journal of Applied Engineering Research, 12(21), 11359-11368.

Tubishat, M., Alswaitti, M., Mirjalili, S., Al-Garadi, M. A., & Rana, T. A. (2020). Dynamic butterfly optimization algorithm for feature selection. IEEE Access, 8, 194303-194314. https://doi.org/10.1109/ACCESS.2020.3033757.

Vashisth, S., Dhall, I., & Saraswat, S. (2020). Chronic kidney disease (CKD) diagnosis using multi-layer perceptron classifier. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 346-350). https://doi.org/10.1109/Confluence47617.2020.9058178.

Yildirim, P. (2017). Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In 41st IEEE Annual Computer Software and Applications Conference (COMPSAC), Vol. 2 (pp. 193-198). https://doi.org/10.1109/COMPSAC.2017.84.

Zhang, Y., Li, X., & Li, Y. (2020). An intelligent chronic kidney disease prediction model based on improved extreme learning machine. Journal of Computational Science, 44, 101196. https://doi.org/10.1016/j.jocs.2020.101196.