Evolving A Neural Network to Predict Diabetic Neuropathy

Shiva Shankar Reddy¹⁺, Gadiraju Mahesh² and N. Meghana Preethi³

¹Research Scholar, Department of CSE, BPUT, Rourkela, Odisha, INDIA
²Associate Professor, Department of CSE, SRKR Engineering College, Bhimavaram, Andhrapradesh, INDIA
³Department of CSE, SRKR Engineering College, Bhimavaram, Andhrapradesh, INDIA

Abstract

One of the main areas where machine learning (ML) techniques are used vastly is in prediction of diseases. Diabetic neuropathy (DN) disease is a complication of diabetes which causes damage to nerves. Early prediction of DN helps diabetic patient to avoid its complications. The main aim of this work is to identify various risk factors of DN and predict it accurately using ML techniques. Radial basis function (RBF) network is an artificial neural network proposed to obtain better results than traditional ML classification techniques. CART, random forest and logistic regression are existing classification techniques considered. Accuracy, recall, f1 score, area under ROC curve (AUC), Matthews correlation coefficient (MCC) and kolmogorov-smirnov (KS) statistic are performance metrics used to evaluate and compare algorithms. From comparative study it was observed that proposed technique RBF network performed better. The performance metric values obtained for RBF network are accuracy-68.18%, recall-0.909, f1 score-0.7407, AUC-0.6405, MCC-0.4082 and KS statistic-0.5417. Accordingly, the use of RBF network while predicting DN gives accurate and better results.

Keywords: Diabetic neuropathy (DN), diabetes, machine learning (ML), artificial neural network (ANN), radial basis function (RBF) network, CART, random forest (RF), logistic regression (LR), accuracy, recall, f1 score, area under ROC curve (AUC), Matthews correlation coefficient (MCC) and kolmogorov-smirnov statistic (KS).

Received on 11 August 2020, accepted on 30 September 2020, published on 26 October 2020

Chronic disease can be treated temporarily by intake of medicines but cannot be cured completely. Diabetes mellitus is a chronic disease that results in high blood glucose levels. In a survey conducted during 2015 to 2019 by Indian government it was identified that about 13.2% of diabetic patients belong to 70-79 years age group. They also observed that the prevalence of diabetes was about 11.8%. [1] Type-1 and 2 and gestational diabetes are the types of diabetes. The diabetes in pregnant woman is called Gestational diabetes. Type-1 and 2 are the most common types of diabetes which leads to severe complications. In India the chance of affecting to type-1diabetes is less than type-2 diabetes. About 1/3rd part of the type-2 diabetic patients in India are obese. [2]

1. Introduction

In recent days percentage of people suffering with health problems is increasing rapidly. Chronic diseases are the most affected diseases to all the age groups. Early diagnosis of such diseases helps a lot, mainly to reduce the risk of affecting to its complications. Diabetic neuropathy is one such chronic complication of diabetes. The main motive of this work is to reduce effort for diabetes patients by predicting diabetic neuropathy. ML has been using widely in medical field for predicting various diseases. So, few ML classification techniques are considered for predicting DN.

*Corresponding author. Email: shiva.shankar591@gmail.com
DN is the chronic diabetes complication that damages nerves in the body due to high blood glucose levels [3]. A person will have more chance of affecting to DN if he has diabetes previously, uncontrolled sugar levels, kidney disease, diabetic retinopathy, hypertension, obesity and habit of smoking. It can further cause complications like digestion problems, loss of leg, toe or foot, urinary tract infections and hypoglycaemia unawareness [4]. Diabetic neuropathy is divided into four types which affects different parts of body as nerves are present throughout the human body. The four types of DN are peripheral, proximal, autonomic and focal neuropathies. [5]

- The nerves leading to hands, arms, legs and feet are damaged in peripheral neuropathy. Approximately 20% and 50% of type-1 and type-2 diabetes patients are affected to diabetic peripheral neuropathy (DPN). [6]
- Muscle weakness is caused by Proximal neuropathy. The muscles in upper part of hips, buttocks and legs are damaged in this type of neuropathy.
- Autonomic neuropathy damages the autonomic nerves system that helps to perform actions like pumping blood to heart, digestion and breathing.
- Focal neuropathy damages only specific nerves. It mostly affects nerves present in the head and sometimes it also affects nerves present in legs and torso. [5]

In a study performed, it was stated that about 50-70% of diabetic patients are suffering with neuropathy [7]. In [8] they have observed that hypertension, obesity, age and duration of diabetes are the major risk factors and smoking as the secondary risk factor of DSPN which is most common form of neuropathy. They have also identified that diabetic retinopathy is one of the possible comorbidity of neuropathy. Similarly from another study performed [9] they have suggested diabetes duration, age, HbA1c, diabetic retinopathy, smoking and BMI as risk factors of DPN.

A study has been performed to identify the parameters which play a major role in identifying the presence of DPN in type-2 diabetes mellitus. These parameters include age, gender, HbA1c value, duration of diabetes, hypertension, and body mass index (BMI) [10]. DN is diagnosed by conducting physical examination which may include symptoms and medical history. Nerve conduction velocity (NCV) and electromyography (EMG) tests are also performed to diagnose DN. NCV test measures the time taken by nerves to transmit the signals. EMG test helps to know how well the muscles respond to the signals given from nerves. [11]

The proposed algorithm in this paper is radial basis function network. The algorithms random forest, logistic regression and CART algorithms are considered as existing algorithms. These four algorithms are implemented in R programming and compared with each other to identify the best performing algorithm in terms of evaluation metrics. The accuracy, recall, f1 score, area under ROC curve, MCC and KS are the metrics used to evaluate the trained models. The best algorithm is the one which obtains better values of metrics.

2. Literature survey

Hasan Mahmud et al. [12] proposed a framework for predicting diabetes using ML algorithms. The pima dataset from UCI repository is considered for the work. The ML algorithms ANN, naive bayes, SVM, logistic regression, decision tree (DT) and random forest were implemented and compared to identify best performing one. Accuracy, sensitivity, specificity, precision and f1-score are metrics considered for evaluating the algorithms using 10-cross validation technique. Among the six algorithms naive bayes has performed better with the value of accuracy as 74%.

Faizan Zafar et al. [13] proposed their work to predict type-2 diabetes efficiently. Pima dataset from UCI ML repository is used to implement the algorithms. KNN, logistic regression, random forest, DT, gaussian naive bayes, gradient boosting, keras neural network and adaboost are the techniques considered for prediction. These are evaluated and compared using f1-score in case of both the raw dataset and pre-processed dataset. The parameter tuning has been considered along with the gradient boosting technique. This technique has outperformed the remaining techniques with value of f1-score as 0.853 in case of pre-processed dataset.

Messa Komi et al. [14] have consider five data mining techniques for early prediction of diabetes. Algorithms namely logistic regression, SVM, Extreme learning machine, Gaussian mixture model and ANN were considered for implementation. Accuracy of the algorithms were compared. The ANN algorithm has obtained a better accuracy of 89% and identified as the best algorithm.

Dinesh Pandey et al. [15] focused on accurate vessel segmentation. The main techniques considered are phase-preserving denoising, maximum entropy incorporating line detection. Based on these techniques a vessel segmentation method was proposed. This proposed method involves the steps namely pre-processing image, identifying thin, thick blood vessels and image post-processing. The identification of thin blood vessels is done by using local phase preserving denoising, local normalization and maximum entropy thresholding. The extraction and binarization of thick vessels is done by maximum entropy thresholding. DRIVE, STARE, CHASE-DB1, HRF are the four datasets chosen to implement the proposed algorithm. They compared the proposed method with the other methods in literature. It was concluded that proposed technique has performed better with accuracy of 0.9623, 0.9444, 0.9494 and 0.9641 for DRIVE, STARE, CHASE-DB1, HRF datasets respectively.

Rafqul Islam et al. [16] proposed a framework to detect depression based on social network data using ML techniques. The data collected from facebook was considered in their work. Decision tree, k-Nearest Neighbor, SVM and ensemble are the techniques considered. The emotional process, temporal process, linguistic style and including all features are the four sets of data considered after feature extraction. When considering these four datasets the decision tree algorithm has performed better than remaining algorithms. The values of f-measure obtained for DT in case of emotional process, linguistic style, temporal process and
including all features are 72%, 72%, 73% and 71% respectively. 

Hu Li et al. [17] used ensemble learning for classifying streaming data. The method is a multi-window based ensemble technique. The datasets namely Elec, Forest, Airlines, Pooker1, Pooker2, Mushroom, Thyroid1, and Thyroid2 are considered. In the proposed method there are three types of windows. They are used to store the newest minority instances, present batch of records and the ensemble classifier. The ensemble technique has a set of latest sub-classifiers and records used to train the each of the sub-classifier. The majority voting was the technique used for class prediction.

Shanshan Chen et al. [18] had performed a study on early screening of DPN. The data of diabetic patients with DPN from 106 in-hospital patients is collected. Additional gait information from a wearable sensor called ear-worn inertial sensor (e-AR) is also added to the clinical data. L.Rhas been used for predicting the risk of having DPN in diabetes patients. The gait data from wearable sensor combined with clinical data has enhanced the capability of clinical data while prediction. The value of c-index has been increased from 0.75 to 0.84 after addition of gait data.

Cut Fiarni et al. [19] considered some data mining techniques to analyze various risk factors and predict diabetes complications. Retinopathy, nephropathy and neuropathy are the complications of diabetes considered and are predicted using risk factors. DT, naive bayes tree and k-means clustering are three techniques performed to analyze the risk factors for each complication. The influential risk factor for neuropathy is females with BMI more than 25. 68% overall accuracy is obtained for the proposed model.

Tahsir et al. [20] used k-means clustering and combination of ANN(ANN) with stratified k-fold cross validation for predicting six complications of type-2 diabetes. They stated that the increase in BMI and blood glucose level results in complications of diabetes. One of the complications of type-2 diabetes is neuropathy. In case of neuropathy the prediction accuracy of respondent males and females are 86.6% and 88.2% respectively for ANN with stratified k-fold cross validation.

Aruna Pavate et al. [21] proposed fuzzy logic technique for predicting risk of affecting to five major complications of diabetes using a dynamic web application. Vision loss, neuropathy, kidney failure, stroke diseases and heart problem are major complications. The data of patient is provided in the web application by selecting any one of the five complications. The accuracy of 92.5% is achieved by using the application.

Arianna Dagliati et al. [22] predicted the complications (retinopathy, neuropathy and nephropathy) of type-2 diabetes using ML techniques. The variables in electronic health record data used for predicting the disease are different for each complication. Random forest, LR with step-wise feature selection, naive bayes and SVM are performed at different times like 3, 5 and 7 years from first admission to hospital. In case of neuropathy 3 years time horizon has obtained better performance values of accuracy, sensitivity, specificity, PPV, NPV, AUC and MCC.

Rahmani Katigari et al. [23] proposed fuzzy expert system to diagnose DN. Seven diagnostic parameters are considered to detect and categorize the severity of DN. The severity of diabetic neuropathy is divided into four categories mild, moderate, severe and absence. Fuzzy expert system is validated using accuracy, sensitivity and specificity measures. It has achieved 93% accuracy, 89% sensitivity and 98% specificity.

Herbert Jelinek et al. [24] detected severity of DN using machine learning technique GBMLS. Graph based ML system (GBMLS) improves the effectiveness of detecting DN. Multi-scale Allen factor (MAF) determines heart rate variability (HRV) from ECG signals. GBMLS with MAF performed better than hybrid bipartite graph formulation (HBGF), cluster-based graph formulation (CBGF), k-means, k-neighbors, random forest, mean shift, birch, DBSCAN, SVM, DT, nearest centroid (NC), ward hierarchical clustering, gaussian naive bayes, multinomial naive bayes (MNB) and bernoulli naive bayes (BNB).

Aruna Pavate and Nazneen Ansari [25] proposed soft computing techniques to predict risk of affecting to type-2 diabetes and its complications. Fuzzy rule-based system and genetic algorithm combined with k-nearest neighbor techniques have been used diabetes prediction and its complications using medical records of 235 patients. Heart diseases, heart stroke, kidney disease, neuropathy and blindness are predicted with the corresponding risk level. The values of accuracy, sensitivity and specificity obtained for GA with KNN are 95.5%, 95.83% and 86.95% respectively and performed best.

Andreja Picon et al. [26] identified the presence of DN by considering uncertainties while predicting. Patients have been classified into four categories (absent, mild, moderate and severe) based on severity of DN using rule based fuzzy expert system. Fuzzy expert system is validated based on area under ROC curve and kappa coefficient value between predicted and actual values, and performed better.

Vincenzo Lagani et al. [27] developed better performing diabetes complication risk assessment models. The models are developed for different Diabetes sideeffects. Diabetes and Complication Control Trial (DCCT) and the Epidemiology of Diabetes Interventions and Complications study (EDIC) data is considered for developing model. The set of parameters for each complication risk assessment model are different. Internal and external validation has been performed on developed models. External validation includes collection of data and dealing with missing values. Concordance index is considered for internal validation.

Cut Fiarni [28] developed a knowledge management system (KMS) for predicting complications of diabetes based on data from social networks. Knowledge management activities, content based reasoning (CBR) and social network model are combined to develop KMS. It enables sharing information between physician and patient through web based system which makes the decision.

Ruhin Kouser et al. [29] developed heart disease prediction system. Case based reasoning (CBR), ANN and RBF techniques has been implemented for the purpose of predicting disease and prescription. Dataset from Cleveland
Heart Disease database have been considered. ANN integrated with CBR is used to diagnose the type of heart disease and obtained 97% accuracy. CBR combined with RBF provided medical prescription by considering the medical prescription of old patients.

In most of the works the basic ML classifications algorithms are considered to predict a disease. The ANN is one of the classification algorithms which can be used for better prediction. There are several types of ANN which are modifications of ANN. The proposed algorithm RBF network is also a type of ANN which was compared with some other ML classification algorithms used in other works. This comparison was done to prove that RBF network will perform better than basic ML classification algorithms.

This paragraph describes content of each section. Section 3 comprises of the proposed research approach to achieve the objectives. Section 4 comprises the working of proposed algorithm RBF network and brief explanation of remaining three algorithms. Section 5 contains the discussion of results obtained after implementing all algorithms using R programming. In this section results are also provided for all the algorithms including the comparison. Conclusion for the work is provided in section 6 followed by the references.

3. Research Approach

3.1. Objectives of work

Early diagnosis or prediction of a disease is very necessary to prevent future complications. In this work the problem of predicting the presence of DN is considered as it is one of the major chronic diabetes complications. The objectives of this work are to

- Identify various risk factors of DN which plays a major role while predicting the disease.
- Predict DN using machine learning techniques.
- Validate trained models using performance metrics.
- Obtain best technique which predicts or diagnoses DN accurately.

Some risk factors of DN are included in the considered dataset which is used for implementation. R programming has been used for implementing all the techniques and evaluating its results. CART, random forest, LR and radial basis function network are the four techniques supposed to use among which RBF network is the proposed technique. Accuracy, recall, f1 score, AUC, MCC and KS statistic are six performance measures considered to evaluate, compare and obtain best technique among all.

3.2. Dataset

Dataset considered is ‘diabetes complications in populations of Iran’ dataset collected from figshare repository [30]. From this dataset the attributes required to predict DN are considered and in addition some risk factors are also included.

The dataset finally contains 15 attributes and 116 instances. Total cholesterol, diabetic retinopathy and smoking are the risk factor attributes included. Total cholesterol is the sum of low density lipoproteins (LDL), high density lipoproteins (HDL) and 20% of triglyceride. Neuropathy is the target attribute that is used to classify the dataset for predicting whether the person is having neuropathy or not. The dataset is a binary classification dataset. Description of all attributes in dataset is provided in table 1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Gender of the diabetic patient. 1 means male and 2 means female.</td>
</tr>
<tr>
<td>Age</td>
<td>Age of the diabetic patient. It is a numerical value.</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index (BMI) is a numerical value calculated using formula:- weight in kg / (height in m)^2.</td>
</tr>
<tr>
<td>Diabetes type</td>
<td>Type of diabetes the patient has been suffering with. 1 means diabetes type-1, 2 means diabetes type-2.</td>
</tr>
<tr>
<td>Diabetes duration</td>
<td>Duration of diabetes that means number of years the patient has been suffering from diabetes.</td>
</tr>
<tr>
<td>A1C value</td>
<td>HbA1C – value of average sugar level in blood for previous 2 to 3 months obtained by performing haemoglobin A1C test.</td>
</tr>
<tr>
<td>Diabetes treatment</td>
<td>Contains categorical values. 1 means oral treatment i.e. through medicine intake, 2 means insulin treatment, 3 means both oral and insulin.</td>
</tr>
<tr>
<td>Total Choles.</td>
<td>It is a numerical value calculated using formula:-</td>
</tr>
<tr>
<td>Statin</td>
<td>Total cholesterol= LDL+HDL+20%TG.</td>
</tr>
<tr>
<td>Dose</td>
<td>Contains values 0, 20, 40, 80. 0 means no need to use statin. Remaining values represents dosage of statin in milligrams per day.</td>
</tr>
<tr>
<td>Systolic BP</td>
<td>Systolic blood pressure is a numerical value that represents the pressure in the blood flow during contraction of heart muscle. Normal range of SBP is ≤120 mmHg.</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>Diastolic blood pressure is a numerical value that represents the pressure in the blood flow in between heart beats. Normal range of DBP is ≤80 mmHg.</td>
</tr>
<tr>
<td>Diabetic Retinopathy (DR)</td>
<td>Contains categorical values. 1 means suffering with DR, 0 means not suffering with DR.</td>
</tr>
<tr>
<td>Smoking</td>
<td>Contains values 0, 20, 40, 80. 0 means no habit of smoking, 1 means patient has habit of smoking.</td>
</tr>
<tr>
<td>Neuropathy</td>
<td>Contains categorical values. 0 means tested negative for neuropathy, 1 means tested positive for neuropathy.</td>
</tr>
</tbody>
</table>
The histogram representation of the dataset provided in Table 1 is given in Figure 1. Each and every attribute is represented using a histogram individually. The blue vertical bars represent distribution of data. Dashed line represents density distribution of data.

From the Histogram of the dataset it is observed that there is a balanced distribution between the non DN and DN instances for the DN variable. But there is an imbalanced distribution between the DR and non DR for the Diabetic Retinopathy (DR) variable and it is not the subject under consideration. But there is some chance of having both the diabetic retinopathy and neuropathy for a diabetic patient. But it was not true that all the diabetic patients will have both the retinopathy and neuropathy. In this work the subject under consideration is DN prediction and the dataset is balanced in the view of DN.

3.3. System architecture

The Figure 2 represents system architecture for proposed system. Initially dataset is loaded and data pre-processing is performed. Data pre-processing involves checking for missing values, splitting the dataset, deal with categorical attributes.
4. Proposed work

This section comprises of details and working of algorithms. The three algorithms CART, random forest and LR are described briefly. Proposed technique RBF network is described in detail.

4.1. CART

Classification and Regression Tree (CART) algorithm uses an impurity measure namely gini index to construct a decision tree. Among all attributes the attribute with lowest gini index is chosen to split the tree. Leaf nodes of the constructed decision tree represent the predicted target values. Gini index could be computed using formula given below. In formula ‘a’ is each attribute in given instance, F1 and F2 represents subset of instances where each subset belongs to a category of ‘a’, c represents the classes in target variable, Pk is probability that F belongs to class k. [31]

\[
Gini_a(F) = \frac{|F_1|}{|F|} Gini(F_1) + \frac{|F_2|}{|F|} Gini(F_2)
\]

\[
\text{where } Gini(F) = 1 - \sum_{k=1}^c P_k^2
\]

4.2. Random forest

Random forest algorithm uses an ensemble technique with voting strategy. Ensemble technique integrates the output of multiple predictive models and outputs the mostly predicted target value. Bootstrap dataset is a subset of original dataset constructed by selecting random samples. Each instance in this bootstrap dataset is used to train a model using DT algorithm which gives predicted target value for that instance. Voting strategy is employed to obtain mostly predicted target value which is the final output predicted by the model. [32]

4.3. Logistic regression

LR algorithm is used for the classification problem. A decision boundary is constructed to classify the given input instances. Decision boundary that lies between [0, 1] is constructed in S shape using a sigmoid function \(h_0(x)\) or \(\sigma(z)\) which is described below. Depending on a threshold value which is fixed between [0, 1] of decision boundary the target value is predicted. The error obtained can be reduced by using a cost function described below where y represents actual value. [33]

\[h_0(x) = \sigma(z) = \frac{1}{1 + e^{-z}}; \text{ where } z = mx + c\]

\[
\text{Cost}(h_0(x), y) = -y \cdot \log \log (h_0(x)) - (1 - y) \cdot \log(1 - h_0(x))
\]

4.4. Radial Basis Function (RBF) Network

RBF network is also called as RBF neural network. It is an ANN that contains only one hidden layer. This hidden layer is also called as feature vector whose dimension will be increased by using radial basis function as activation function. Each neuron of hidden layer has n-dimensions where n is number of predictor attributes. If the data is linearly not separable then increase in dimension of feature vector makes the data linearly separable. Gaussian radial function is used in hidden layer as activation function. Weights are initialized between hidden and output layers. Output layer contains a neuron for each target class. The weighted sum of the outputs from hidden layer is forwarded to neurons in output layer. Classification is done at the output layer only.

Table 2. Radial Basis Function Network Algorithm

<table>
<thead>
<tr>
<th>Algorithm. Radial Basis Function (RBF) Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: The instances in the dataset.</td>
</tr>
<tr>
<td>Output: Predicted output values of the target attribute.</td>
</tr>
<tr>
<td>Assumptions: x is value of neuron input layer which is connected with neuron in hidden layer, (c_i) is centre of neuron ‘i’ in hidden layer, (c_i) is width of neuron i in hidden layer, n is number of neurons in hidden layer, (W_k) is weight of connections between neurons ‘h’ in hidden layer and ‘k’ in output layer.</td>
</tr>
<tr>
<td>Step-1: Start</td>
</tr>
<tr>
<td>Step-2: Set values of neurons in input layer. Each neuron holds value of a predictor attribute in input instances.</td>
</tr>
<tr>
<td>Step-3: Initialize centre ((c_i)), width ((c_i)) of each neuron ‘i’ in hidden layer and weight ((W_k)) of connections between neurons in hidden and output layer.</td>
</tr>
<tr>
<td>Step-4: For each neuron ‘i’ in hidden layer perform activation function namely Gaussian radial function.</td>
</tr>
<tr>
<td>(\Phi_i(x) = \exp \left(-\frac{r_i^2}{\sigma_i^2}\right); \text{ where } r_i =</td>
</tr>
<tr>
<td>Step-5: Calculate the values of neurons in output layer using below formula</td>
</tr>
<tr>
<td>(F_k(x) = \sum_{h=1}^{n} W_{hk} \cdot \Phi_h(x))</td>
</tr>
<tr>
<td>Step-6: The highest value among all neurons in output layer is given as output.</td>
</tr>
<tr>
<td>Step-7: Stop</td>
</tr>
</tbody>
</table>

Table 2 is shows the algorithm for the proposed algorithm RBF network. The values are given for each input layer neuron in step-2. Number of neurons in input layer is equal to number of predictor attributes. Initializing the values for centre, width of hidden layer neurons and weight of connections between hidden and output layers is done in step-3. The values of centre for all hidden neurons are assigned by using k-means clustering algorithm. Values of width and weights of connections are assigned by using error back propagation which is a supervised training process. In step-4, activation function called Gaussian radial function is
performed for each hidden layer neuron. These values are used to calculate values of neurons in output layer. In step-5 weighted sum of the values obtained in step-4 is given to output neurons. In step-6 the output values are obtained. For classification the output layer neurons count is equal to target attributes classes or categories count. The neuron which obtained highest value is given as final predicted value. \[34\]

The input layer contains 14 neurons one for each predictor attribute. The no. of neurons in hidden layer is considered as 20. The parameters in the algorithm like weights between hidden and output layer, centre and width of the hidden neurons are initialized by implementing the RBF network using a built-in function in R programming namely ‘rbf’. The output layer contains 2 neurons as there are two possible classes for the target attribute. One neuron holds the value for positive class and the other will hold the value for the negative class. After obtaining the values for output neurons as mentioned in table 2. The class to which the value obtained is higher is the predicted value of the target attribute.

5. Results & Discussions

This section comprises of results obtained after implementing all algorithms and comparison between them. Results of each algorithm are considered and compared in terms of accuracy, recall, f1-score, AUC, MCC and KS statistic. All results are obtained by implementing algorithms using R programming. The evaluation of all considered performance metrics is provided for proposed algorithm RBF network. Similarly remaining algorithms are also evaluated. The values of TP, FP, TN and FN in confusion matrix obtained for RBF network are 10, 6, 5 and 1 respectively. Using these values the performance metrics of RBF network are calculated below.

5.1. Performance metrics

Accuracy

Accuracy is the number of records correctly predicted out of total number of records.

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{10+5}{22} = 0.6818
\]

Recall

Recall is the ratio of the number of records that were correctly predicted as positive to total records which should be identified as positive.

\[
\text{Recall} = \frac{TP}{TP+FN} = \frac{10}{11} = 0.9090
\]

F1 Score

F1 Score is the harmonic mean between precision and recall. Precision is the number of instances which are correctly predicted as positive divided by total positive predicted instances.

\[
\text{Precision} = \frac{TP}{TP+FP} = \frac{10}{16} = 0.625
\]

\[
\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.625 \times 0.9090}{0.625 + 0.9090} = 0.7407
\]

Area under ROC Curve

AUC is based on ROC curve. ROC curve is the graph plotted between recall and 1-specificity. Its range is 0 to 1. Value nearer 1 indicates better prediction. AUC value obtained by plotting ROC curve for RBF network result is 0.6405.

\[
1 - \text{Specificity} = 1 - \frac{TN}{TN+FP} = 1 - \frac{5}{11} = 0.5455
\]

Matthews’s correlation coefficient

MCC measures the quality of binary classification. It is correlation between predicted classes and actual classes that is calculated based on values obtained in confusion matrix. Its value lies between -1 and 1. Value nearer to 1 indicates better prediction.

\[
\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}
\]

\[
= \frac{(10 \times 5) - (6 \times 1)}{\sqrt{(10+6)(10+1)(5+6)(5+1)}} = \frac{44}{107.77} = 0.4082
\]

Kolmogorov-Smirnov statistic

KS statistic is the maximum difference between cumulative percentage of true positive rate and cumulative percentage of false positive rate. The instances in the dataset are divided into 10 equal sized bins and then the maximum value of ks-statistic obtained in these bins is given as output. Higher value indicates better prediction. KS statistic value obtained for RBF network is 0.5417.

5.2. Results obtained

CART

\[\text{Figure 3. ROC curve obtained for CART}\]
Table 3. Results obtained for CART

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
<th>MCC</th>
<th>KS-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>63.64%</td>
<td>0.5455</td>
<td>0.6</td>
<td>0.5385</td>
<td>0.2773</td>
<td>0.2393</td>
</tr>
</tbody>
</table>

The evaluation of the performance metrics is done in the same way as RBF network which is provided above. Values of six performance metrics accuracy, recall, F1 score, AUC, MCC and KS statistic obtained for CART is provided in table 3. Value of AUC is obtained from ROC curve of CART in figure 3.

Random forest

The evaluation of the performance metrics is done in the same way as RBF network. Values of six performance metrics accuracy, recall, F1 score, AUC, MCC and KS statistic obtained for random forest is provided in table 4. The Value of AUC is obtained from ROC curve of random forest in figure 4.

Table 4. Results obtained for Random forest

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
<th>MCC</th>
<th>KS-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.09%</td>
<td>0.5455</td>
<td>0.5714</td>
<td>0.475</td>
<td>0.1825</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Logistic regression

The evaluation of the performance metrics is done in the same way as RBF network. Values of six performance metrics accuracy, recall, F1 score, AUC, MCC and KS statistic obtained for LR is provided in table 5. Value of AUC is obtained from ROC curve of LR in figure 5.

Table 5. Results obtained for Logistic regression

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
<th>MCC</th>
<th>KS-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.18%</td>
<td>0.8182</td>
<td>0.72</td>
<td>0.5625</td>
<td>0.3779</td>
<td>0.3214</td>
</tr>
</tbody>
</table>

Radial Basis Function Network

Values of six performance metrics accuracy, recall, F1 score, AUC, MCC and KS statistic obtained for radial basis function network is provided in table 6. Value of AUC is obtained from ROC curve of RBF network in figure 6.

Table 6. Results obtained for RBF network

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
<th>MCC</th>
<th>KS-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.18%</td>
<td>0.909</td>
<td>0.7407</td>
<td>0.6405</td>
<td>0.4082</td>
<td>0.5417</td>
</tr>
</tbody>
</table>

Figure 4. ROC curve obtained for Random forest

Figure 5. ROC curve obtained for Logistic regression

Figure 6. ROC curve obtained for RBF network
Comparison of algorithms

Results obtained for all the algorithms are given in table 7. It is visualized that the values of six performance metrics accuracy, recall, F1 score, AUC, MCC and KS statistic obtained for RBF network are better than remaining algorithms values. Evaluation of all the performance metrics for each algorithm is performed but provided only for RBF network. By comparing all these metrics the best algorithm which obtained better values in each metric is obtained. From above result analysis it was observed that RBF network has performed better with values of accuracy, recall, F1 score, AUC, MCC and KS statistic as 68.18%, 0.909, 0.7407, 0.6405, 0.4082 and 0.5417 respectively. Since RBF network has performed better it was identified as best algorithm.

Table 8 comprises of the analysis of the proposed work and other ML techniques from literature survey in section 2. In literature survey, works are done on different complications of diabetes which included neuropathy using few ML techniques. While in some other works only a single ML algorithm is proposed but not compared with existing algorithms.

Generally ANN is preferred while solving or predicting complex problems. So in this work RBF network is proposed to predict DN. RBF network is a type of ANN that contains only a single hidden layer. So, the training process of RBF network is faster than traditional ANN which may contain more than one hidden layer. In this work the comparison of RBF network is performed with three ML classification techniques like random forest, LR and CART decision tree.

Table 7. Results of all algorithms for comparison

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
<th>MCC</th>
<th>KS-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>63.64%</td>
<td>0.5455</td>
<td>0.6</td>
<td>0.5385</td>
<td>0.2773</td>
<td>0.2393</td>
</tr>
<tr>
<td>Random forest</td>
<td>59.09%</td>
<td>0.5455</td>
<td>0.5714</td>
<td>0.475</td>
<td>0.1825</td>
<td>0.15</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>68.18%</td>
<td>0.8182</td>
<td>0.72</td>
<td>0.5625</td>
<td>0.3779</td>
<td>0.3214</td>
</tr>
<tr>
<td>RBF network</td>
<td>68.18%</td>
<td>0.9090</td>
<td>0.7407</td>
<td>0.6405</td>
<td>0.4082</td>
<td>0.5417</td>
</tr>
</tbody>
</table>

Table 8. Analysis of proposed work and ML techniques from literature survey

<table>
<thead>
<tr>
<th>Author</th>
<th>Techniques considered</th>
<th>Findings in the work</th>
<th>Best technique</th>
<th>Values of metrics for best technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>CART, random forest, logistic regression and RBF network</td>
<td>Identify various risk factors of DN and predict it using ML techniques. The RBF network is the proposed technique compared with three other techniques. Accuracy, recall, F1 score, AUC, MCC and KS statistic are evaluation metrics based on which comparison is done.</td>
<td>RBF network</td>
<td>Accuracy-68.18%, recall-0.909, F1 score-0.7407, AUC-0.6405, MCC-0.4082 and KS statistic-0.5417</td>
</tr>
<tr>
<td>Shanshan Chen et al. [18]</td>
<td>Logistic regression</td>
<td>Performed a case study on diagnosis of diabetic peripheral neuropathy in diabetic patients. Gait data from a wearable ear sensor is added to clinical data. C-index is considered for model evaluation.</td>
<td>Logistic regression</td>
<td>The value of C-index obtained was 0.84 when gait data is added to clinical data. Value obtained is 0.75 when only MNSI clinical history dataset is used.</td>
</tr>
<tr>
<td>Authors</td>
<td>Techniques</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiarni et al. [19]</td>
<td>C4.5 decision tree, naive bayes (NB) tree and k-means clustering</td>
<td>Data mining techniques are used to identify most influential risk factors for three complications of diabetes and predict them. They are retinopathy, nephropathy and neuropathy. Accuracy is the metric chosen for evaluating the final proposed model. This means overall average accuracy considering three complications. Rather than best technique the authors identified the complication which has obtained highest prediction accuracy, which is retinopathy.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tahsir et al. [20]</td>
<td>ANN with stratified k-fold cross validation and k-means clustering</td>
<td>Used the techniques for predicting six complications of type-2 diabetes which also include neuropathy. Accuracy is the only metric used to evaluate and predict the complications. ANN with stratified k-fold cross validation Accuracy values obtained for male and female respondents in case of neuropathy are 86.6% and 88.2% respectively.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aruna Pavate et al. [21]</td>
<td>Fuzzy logic</td>
<td>Fuzzy logic is used to develop a dynamic web application to predict the risk level of five diabetes complications. Neuropathy is one of the five complications they have considered. Performance of the model is evaluated using accuracy.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arianna Dagliati et al. [22]</td>
<td>Random forest, logistic regression with step-wise feature selection, naive bayes and SVM</td>
<td>Complications of type-2 diabetes are predicted using ML techniques. These complications include retinopathy, nephropathy and neuropathy. AUC is used to identify best algorithm. Further the best algorithm is evaluated using some other metrics for 3.5 and 7 years time horizon after first admission to hospital. All the metrics considered for best technique are accuracy, sensitivity, specificity, PPV, NPV, AUC and MCC. LR with step-wise feature selection has performed better. In case of neuropathy, 3 years time horizon has given better values than 5 and 7 years. The values are Accuracy-0.746, sensitivity-0.783, specificity-0.707, PPV-0.743, NPV-0.750, MCC-0.490, and AUC-0.799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rahmani Katigari et al. [23]</td>
<td>Fuzzy expert system</td>
<td>Diagnosed the DN severity using fuzzy expert system. Accuracy, sensitivity and specificity evaluation metrics were chosen for evaluation of model. Fuzzy expert system Accuracy-93%, sensitivity-89% and specificity-98%.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbert Jelinek et al. [24]</td>
<td>Graph based machine learning system with multi-scale allen factor (MAF), HBGF, CBGF, k-means, k-neighbors, random forest, mean shift, birch, DBSCAN, SVM, decision tree, nearest centroid (NC), ward hierarchical clustering, GNB, MNB and BNB</td>
<td>Detection of DN is performed using several ML techniques. The proposed technique is GBMLS with MAF. MAF technique will determine the heart rate variability (HRV) from ECG signals. The sensitivity and specificity are the metrics they have considered for evaluating and comparing all the models. GBMLS with MAF Sensitivity-0.89 and specificity-0.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the proposed work, firstly the risk factors related to neuropathy were identified and those are included in the dataset to predict neuropathy. By including the risk factors the prediction of the disease will be efficient. As most of researchers used traditional ML techniques, a type of ANN namely RBF network was compared with some traditional ML techniques in this work. In case of the dataset considered, the RBF network has performed better than other three traditional ML techniques that means it has outperformed the three existing algorithms. This proves that including the risk factor of a disease and using neural network other than ANN like RBF network can give better results for DN.

In this work, though accuracy obtained was not a best value, the values obtained for other metrics were better. On the basis of those performance metrics the RBF network is identified as the best performing algorithm.

Accordingly, using RBF network for prediction of DN will give good results.

6. Conclusion

This work is mainly focused on predicting one of the chronic complications of diabetes namely diabetic neuropathy. Most of the people in this world are affected to diabetes. In this scenario predicting DN in early stage is very necessary to avoid further complications. ML technique namely radial basis function network is proposed to use for prediction purpose. CART, random forest and LR are some existing traditional classification algorithms which are also implemented and compared with RBF network. From comparative study performed, RBF network has achieved good results with values of accuracy, recall, f1 score, AUC, MCC and KS statistic as 68.18%, 0.909, 0.7407, 0.6405, 0.4082 and 0.5417 respectively. The severity of neuropathy is classified into 4 categories (absent, mild, moderate, and severe). This is done using fuzzy expert system. ROC curve area is considered for evaluating the model. Kappa statistic is used for agreement analysis between expert classification and model.

| Aruna Pavate and Nazneen Ansari [25] | Genetic algorithm with KNN and fuzzy rule based system | Used soft computing techniques to predict type-2 diabetes complications. Neuropathy is considered as one among the complications. GA with KNN is used to select best feature subset and predict the disease. To check further complications fuzzy rule based system is used to predict its risk level. Accuracy, sensitivity and specificity are the metrics considered. | They considered prediction of a single disease and provided results for GA with KNN which performed best. | Accuracy-95.5%, sensitivity-95.83% and specificity-86.95%. |
| Andreja Picon et al. [26] | Rule based fuzzy expert system | The severity of neuropathy is classified into 4 categories (absent, mild, moderate, and severe). This is done using fuzzy expert system. ROC curve area is considered for evaluating the model. Kappa statistic is used for agreement analysis between expert classification and model. | Rule based fuzzy expert system | ROC curve area-0.91 and from the agreement analysis using kappa statistic they stated that the model and experts agree with each other. |

References


