Received: June 1, 2019

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Feature Selection Effects on Gradient Descent Logistic Regression for Medical Data Classification

Prasanth Kumar[1]* **Gera Pradeepini[1]** **Pille Kamakshi[2]**

[1]*Koneru Lakshmaiah Deemed to Be University, Andhra Pradesh, India*
[2]*Kakatiya Institute of Technology and Science, Andhra Pradesh, India*
* Corresponding author's Email: prashanth.17883@gmail.com

Abstract: In recent years, a number of researchers have concentrated on medical data analytics because machine intelligence in medical diagnosis is a new trend for enormous medical applications. Generally, medical datasets are massive in size, so traditional classifiers suffered from overfitting and under-fitting problem of training set. In this paper, Gradient Descent Logistic Regression (GDLR) classification method is proposed for medical data classification. The Pearson Correlation Coefficient (PCC) is used to calculate the correlation between the features. After that, Random Forest (RF) algorithm ranks the features and selects the most relevant features to improve performance of the medical data classification. The regression technique processes the features effective and analyse the feature importance based on the weight values. The Random Forest (RF) assigns the features importance in the tree structure. The random forest is used to select the features and features are applied for the GDLR to classify effectively. The GDLR method further analysis the features for effectively analysis the feature importance based on the weight values and more relevant features are identified than the RF. The experimental analysis demonstrated that the performance of GDLR algorithm achieved better than traditional methods Neural Network for Threshold Selection (NNTS) and Mean Selection (MS). The accuracy of the proposed GDLR method achieved as 97.5% in the Hepatitis dataset, while existing mean selection method has the accuracy of 82.58%.

Keywords: Gradient descent logistic regression classification, Mean selection, Neural network for threshold selection, Pearson correlation coefficient.

1. Introduction

Medical data mining has become an emerging area in data mining in recent years and many researchers' utilized different tools for developing medical expert systems. Presently, medical dataset is increasing enormously day by day, numerous data mining techniques emerged to handle large scale of datasets. The data mining technique is employed in different applications such as e-business, web mining, data prediction, medicine analysis, etc. Compare to other fields, medical database management system generated a number of medical databases, hence volume of medical data increased day-by-day [1, 2]. The medical diagnosis and prognosis process are generally having complexity and uncertainty problems in making decisions [3]. Furthermore, a new system is used to improve the diagnosis and prognosis accuracy, namely Clinical Decision Support System (CDSS) [4]. Many researchers developed various machine learning and data analysis methods for medical data clustering [5], classification [6], diagnosing different diseases etc. [7]. Machine learning models developed to support various medical decision making tasks. As an example, intelligent classifiers are in use for prognosis, diagnosis, screening of diabetes, breast cancer and Parkinson disease from the UCI database [8].

In healthcare applications, classification analysis has been significantly implemented to support medical decisions, prognosis, etc. [9]. A traditional medical data analysis models still lag in classification of medical datasets. Several problems adversely affect the classification performance, such as curse of

dimensionality, incomplete dataset [10], irrelevant or redundant features decreases the classification accuracy and increases the computation time [11]. The major contribution of this research work is to provide an efficient medical data classification, so the GDLR classification method is proposed. It is performed on multiple medical datasets Cleveland, Hepatitis, Pima Indians Diabetes (PID), and Wisconsin Breast Cancer (WBC). The raw input data taken from the medical datasets, but those datasets are in different ranges. It is possible that the performance of classification will degrade due to different ranges of multiple datasets, so to rectify this problem normalization method is applied. Next, Pearson Correlation Coefficient (PCC) method calculates the presence or absence of correlation between the two variables and determines the exact level of the correlation. These correlated feature values rank by Random Forest (RF) algorithm and the highly correlated features are given as input to the classifier. The selected correlated features improve performance of the classification. Finally, GDLR classifier efficiently predicts the normal and abnormal data with respect to different medical datasets. The major benefit of GDLR classifier is that it can easily update the new data to classification model with the help of Gradient Descent (GD) method and handle larger training sets. The GDLR method is proposed in this method for data classification to increase the performance of classification. The GDLR has the advantages of the require lower computation resources, high interpretable, avoid overfitting regularly and doesn't require input features to be scaled. Therefore, the proposed GDLR provides the efficient performance in the classification. The data easily understand by the GDLR method that eliminates the overfitting problem.

This paper is composed as follows. Section II presents survey of several recent papers on medical data classification strategies. In section III, an effective feature selection and classification method PCC, RF and GDLR approach is presented. Section IV shows comparative experimental result of proposed and existing classification strategies using UCI medical dataset. The conclusion is made in section V.

2. Related work

Researchers in medical data classification have suggested several research techniques. A brief evaluation of some essential contributions to the existing literatures presented in this section.

M. Seera, C. P. Lim, S. C. Tan, and C. K. Loo, [12] presented a hybrid model for data classification, namely Fuzzy ARTMAP (FAM) neural network with classification regression tree (CART). The major benefit of FAM method is stability-plasticity dilemma that affects the data based learning system. The CART method explicitly represents the learned knowledge in a tree structure. The FAM-CART method able to learn knowledge from the samples and extract the useful rules by rectifying the significant problems. An experimental analysis demonstrated that proposed FAM-CART method employed in six benchmark UCI datasets. This system is unable to handle noisy data, which is the main drawback.

Y. Xu, [13] presented efficient imbalanced data classification using Maximum Margin of Twin Spheres Support Vector Machine (MMTSSVM). This method initiates two homocentric spheres through solving a smaller-sized QPP and Linear Programming Problem (LPP). The small sphere captured many samples in majority class and the large sphere captured most samples in minority class. The margin between the two homocentric spheres maximized and the method worked faster than the other Twin Support Vector Machine (TSVM) based models. The experimental results showed that the MMTSSVM is feasible and valid. The MMTSSVM method degrades the classification performance due to irrelevant features, so a suitable feature selection algorithm is required.

S. Yang, J. Z. Guo, and J. W. Jin, [14] proposed Improved Iterative Dichotomiser 3 (IID3) algorithm for disease prediction. The proposed IID3 algorithm includes three features, (i) decrease the weight of attributes by balance function, (ii) discretization algorithm avoids the manual section of optimal partition numbers, and (iii) rule based heuristic strategy to decrease the memory usage. The experimental result of IID3 algorithm was compared to the traditional classifiers such as random tree, decision stump, in terms of accuracy, stability and minor error rate. However, IID3 algorithm lacks in meeting the big data computational demand.

L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, [15] proposed k-Nearest Neighbor (KNN) classifier with distance functions for medical data classification. The distance function improves accuracy of the classification and calculate the distance between the test data and every training data. According to the experimental analysis, three various kinds of datasets were used, those are, numerical, categorical and mixed kinds of data with four distinct distance functions cosine, Euclidean, chi square and Minkowsky. Here, Euclidean, cosine and Minkowsky distance measures failed to perform over

the mixed type of datasets that affects the classification accuracy of the KNN classifier.

L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, [16] presented SVM with Fruit Fly Optimization Algorithm (FOA) for medical diagnosis. The FOA-SVM method increase the generalization capacity of the SVM classifier by swarm intelligence technique for optimal parameter tuning. The experimental results showed that the FOA-SVM achieved average CPU time cost, so there is a need of alternative medical decision support technique.

Jaganathan and Kuppuchamy, [17] developed the fuzzy entropy based on the feature relevance and analyzed the method Radial Basis Function Network Classifier for a medical databases classification. Three feature selections are used for the medical data classification. This method is tested with different dataset and shows that the method can classify the data with considerable performance. The efficiency of the method is need to be developed by using suitable feature selection technique.

S.M.S. Shah, S. Batool, I. Khan, M.U. Ashraf, S.H. Abbas, and S.A. Hussain [18] proposed Probabilistic Principal Component Analysis (PPCA) technique for the missing data attributes in the medical data classification. PPCA method extracts the feature vectors that contain highest covariance and used for the feature selection. The data has been classified with the help of the Support Vector Machine (SVM). The efficiency of the method is low and can apply feature selection technique to increase performance.

S. Bashir [19] presented heterogeneous classifier namely HMV for medical data classification. This classifier resolved the storage problem by selecting the important features for disease analysis, but it is difficult to perform in imbalanced medical datasets.

So, Gradient Descent Logistic Regression (GDLR) classification method is implemented to overcome the above-mentioned drawbacks and for enhancing the recognition rate of normal and abnormal prediction of medical data.

3. Proposed Methodology

The GDLR based medical data classification approach has four major steps; those are medical data acquisition, preprocessing, feature selection and classification. Initially, the medical data are acquired from the different UCI machine learning repository dataset Cleveland, Hepatitis, Pima Indians Diabetes, and Wisconsin Breast Cancer. In Second stage, preprocessing step is carried out by min-max normalization method because all datasets belong to different ranges, so it is converted to [0, 1] range. Next, PCC method is used for calculating the correlation between the features. The RF algorithm ranks these features. This algorithm selects the highly correlated values and forward to the classifier. Finally, classification process is carried out by GDLR method that predict the normal and abnormal data. The graphical representation of the proposed block diagram is shown in the Fig. 1.

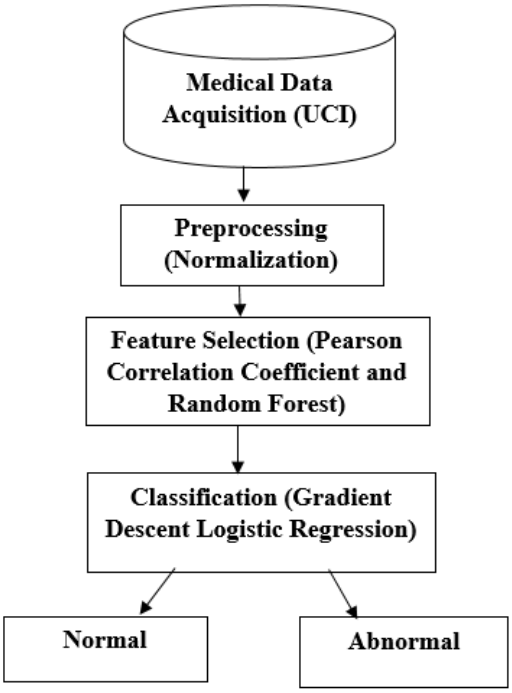

Figure.1 General block diagram of proposed medical data classification

3.1 Medical data acquisition

In experimental analysis, different types of medical databases Cleveland, PID, WBC, Hepatitis databases available in the UCI machine-learning repository are used.

3.1.1. Cleveland data

The Cleveland database comprise of 76 features and all data samples are listed under 14 categories [20]. It represents the value from 0 to 4 stages of heart disease. For privacy purpose, some of the patient's name and their social security number are removed from the data and replaced with the dummy values. The six features are removed from 76 features due to incomplete tests performed among the values. This database consists of 56% of sample heart disease data and 46% are not belongs to heart disease.

3.1.2. Hepatitis data

This dataset is obtained from the Carnegie-Mellon University and it includes 155 instances belongs two classes such as absence or presence. This dataset includes several features such as sex, age, fatigue, liver film etc. [21].

3.1.3. Pima Indians Diabetes (PID)

PID dataset stands for Pima Indians Diabetes. This database comprises of eight attributes and 768 instances, from National Institute of Diabetes, Digestive and Kidney disease [22]. In this dataset, 0 value determine negative result and 1 indicates a positive result.

3.1.4. Wisconsin Breast Cancer (WBC)

This database is collected from the UCI machine-learning repository [23]. The WBC dataset is collected by Dr. Willian H. Wolberg in the year of 1989-1991. The WBC dataset includes 699 samples, which are categorized by different features such as cell size, shape, bland chromatin, benign or malignant growth, etc. In this dataset, approximately 34.5% of (241 samples) malignant and approximately 65.5% of (458 samples) instances are benign.

The input data is taken from these four databases that comprise of different kinds of diseases like heart disease, diabetes disease, mammogram related diseases, etc. These raw data are forwarded to the preprocessing step.

3.2 Pre-processing

The preprocessing step fill the missing value, identify or remove the outliers and resolve inconsistencies of data. The raw data have some noise or errors; it is very important to mine the data in order to get better outcomes from the given data set. In this research work, min-max normalization method is used for preprocessing because all the attributes are different ranges in the dataset, so it is converted into [0, 1] range.

The min-max method is the one kind of normalization technique and it standardize the dataset using linear transformation. This normalization method transforms the medical input data into fixed range. Min-Max method preserves the associations between the original input value and the scaled value. In addition, an out of bound error is encountered when the normalized values deviate from the original data range. This technique ensures that extreme input values are constrained within a specific range. Min-max normalization transforms a value X_0 to X_n which fits in the specified range and it is given by the Eq. (1),

$$X_n = \frac{X_0 - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Whereas, X_n is a new value for variable X, X_0 is a current value for variable X, $and\ X_{min}$ is the minimum data point in the dataset and X_{max} is the maximum data point in the dataset. The min-max normalization method maps a value between X_0 to X_n in the range$[0,1]$. Hence, assume that $X_{min} = 0$ and $X_{max} = 1$.

3.3 Feature selection

Feature selection is the significant process because it improves the classification accuracy performance. It selects a small subset of features from the original feature space. As a result, it avoids noisy data, redundancy and selects the most relevant features for medical data classification. Generally, the feature selection method is categorized into two approaches (i) filter based method and (ii) wrapper method. The filer-based approach depends on the general features of the data to choose the new feature subset for example, chi-square, information gain, etc. The wrapper approach employs predetermined machine-learning algorithm to select the new subset of features, for example, Genetic Algorithm, Bayesian Network, etc. [24]. In this research work, PCC is used to calculate the features and Random Forest algorithm is used for ranking the features.

The PCC method is also known as linear correlation and it calculates the similarity measure between two random variables. The PCC represented as ρ and it estimate the dependency between the two random variables X and Y. Consider pairing of variables X with value x_i and Y with value y_i. The estimation of PCC is given in Eq. (2),

$$\rho = \frac{cov(X,Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}} \quad (2)$$

Where, cov is the covariance and σ is the variance. The mathematical definition of PCC is represented in Eq. (3),

$$\rho = \frac{\sum_i (x_i - \bar{x_i})(y_i - \bar{y_i})}{\sqrt{\sum_i (x_i - \bar{x_i})^2 \sum_i (y_i - \bar{y})^2}} \quad (3)$$

Whereas, x_i is the mean of X, and y_i is the mean of Y. The value of ρ lies between -1 and 1, if X and Y are linearly dependent (correlated), and $\rho = 0$ if X and Y are totally independent (uncorrelated).

The RF learning algorithm helps to minimize prediction error and selects the most relevant features for classification. Random forests construct a collection of trees, where each tree is grown by random independent data sampling & feature splitting, produce a collection of independent identically distributed trees. This algorithm provides the highly correlated features to the GDLR classifier for improving the medical data classification. The goal of the random forest function F minimizes the expected loss subject in training set is shown in Eq. (4).

$$\min_{f \in F} E_{xy}[L(y, f(x))], \text{ s.t. } E_x[C(f, x)] \leq B, \quad (4)$$

Where $L(y, \hat{y})$ is a loss function, $C(f, x)$ is the cost of evaluating the function of F on example x and B is a user specified budget constraint. The feature acquisition cost $C(f, x)$ is a modular function helps to calculate the individual feature cost. The $C(f, x)$ is used by function f on example x, to calculate the cost of each samples. Then minimize the empirical loss subject to a budget constraint, which is shown in the Eq. (5),

$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)), \text{ s.t. } \frac{1}{n} \sum_{i=1}^{n} C(f, x_i) \leq B \quad (5)$$

In our context the classifier f is a random forest, T consisting of K random trees, $D1, D2, \ldots, DK$ are learnt on training data. Consequently, the expected cost for an instance x during prediction-time is written in Eq. (6),

$$E_f\left[E_x[C(f, x)]\right] \leq \sum_{j=1}^{K} E_{Dj}\left[E_x\left[C\left(D_j, x\right)\right]\right] \quad (6)$$

The random forest tree algorithm equally distributes the RHS scale value with the number of trees. The upper bound of the trees show the typical behavior of the random forest because of low features correlation between the trees. With the help of RF approach, classifiers spend less time to classify the normal and abnormal data from different medical datasets. The features redundancies are detected by correlation analysis and improves the data classification performance.

3.4 Classification

After obtaining the optimal feature information, the classification is performed on the extracted data. Classification is defined as a boundary between the classes in order to label the classes based on their measured features.

The logistic regression model is a machine-learning model used to predict the probability of occurrence of an event by fitting data to a logistic curve. This model is used in different applications such as biomedicine, social science, genetics, etc. Let's assume that learning data pairs are represented as (x_i, y_i) of a vector of co-variates indicted as $x_i = (x_1, \ldots . x_n) \in \mathbb{R}^n$ and dependent variable is represented as $y_i \in \{\pm 1\}$. At first, logistic regression model finds the maximum likelihood estimation of the optimal value is indicated as $\beta \in \mathbb{R}^{n+1}$ which increases the probability value. The Eq. (7) represents the estimation of maximum likelihood using logistic regression model.

$$\prod_{i=1}^{n} Pr(y_i | x_i) = \prod_{i=1}^{n} \frac{1}{1 + exp(-y_i (1, x_i)^T \beta)} \quad (7)$$

A logistic regression model consists of only original covariates but, the model includes nonlinear covariates on that condition overfitting of training set problem is elevated. Hence, it decreases the model prediction accuracy. To overcome this issue, minimized loss function is used which defined negative log-probability is shown in Eq. (8),

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} log\left(1 + exp\left(-z_i^T \beta\right)\right) \quad (8)$$

Whereas, $z_i = y_i . (1, x_i)$ for $i = 1, \ldots n$ and z_i is the training set of n samples. If an actual label of particular data point $y_i = 0$ and the predicted probability of $x_i = 1$, then increases the cost function of the logistic function. Likewise, if data points of x_i and y_i are same then the cost function is zero. Hence, finding a minimum cost function is an essential task, so the gradient descent algorithm is used. The Eq. (9) represents the minimized loss function.

$$\nabla J(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \sigma\left(-z_i^T \beta\right) . z_i \quad (9)$$

Whereas, $(x) = \frac{1}{1 + \exp(-1)}$. The descent method starting from initial point β_0 and each step represented as t updates the regression parameters using the Eq. (10),

$$\beta^{(t+1)} = \beta^t + \frac{\alpha_t}{n} \sum_{i=1}^{n} \sigma\left(-z_i^T \beta^{(t)}\right) . z_i \quad (10)$$

Whereas, α_t is a learning rate at step t. This method finds a local minimum or maximum of a function by moving along gradients. To minimize the function in the direction of the gradient, one-

dimensional optimization methods used here. As a result, the GD method decreases the error rate and improves the classification accuracy. At the last stage, GDLR classifier predicts the medical data as two categories such as normal data and abnormal data.

4. Experimental result and discussion

For experimental simulation, Python 3.6.5 JupyterLab software was employed in the PC with 3.2 GHz and i5 processor. The performance of the GDLR is compared to the traditional methods such as PSO+ Extreme Learning Machine (ELM), and Neural Network Threshold Selection (NNTS) [17, 18]. The performance of the proposed classifier GDLR was evaluated by means of accuracy, sensitivity and specificity.

4.1 Performance measure

Performance measure is defined as the relationship between the input and output variables of a system understand by employing suitable performance metrics like sensitivity and specificity. The general formula for calculating the specificity and sensitivity of the pedestrian data is given in the Eq. (11) and (12).

$$Specificity = \frac{TN}{TN+FP} \tag{11}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{12}$$

Accuracy is the measure of statistical variability and a description of random errors. The general formula of calculating pedestrian data accuracy performance is given in the Eq. (13).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{13}$$

Where, TP is represented as true positive, FP is denoted as false negative, TN is represented as true negative and FN is stated as a false negative.

4.2 Performance analysis using different UCI datasets

In this section, UCI database is used for comparing the performance of the proposed GDLR method and the existing methods MS and NNTS [17]. In Table 1, the performance of proposed and existing methods is validated by means of accuracy, sensitivity and specificity. The following Table 1 shows the result of medical data classification performance of proposed and existing method with respect to four different datasets Cleveland, PID, WBC, and Hepatitis.

Table 1. Performance evaluation of different datasets

Dataset	**Mean Selection [17]**		
	Accuracy	**Specificity**	**Sensitivity**
Cleveland	81.75	82	82
PID	76.04	78	71
WBC	95.99	97	93
Hepatitis	82.58	60	87
Dataset	**NNTS [17]**		
	Accuracy	**Specificity**	**Sensitivity**
Cleveland	84.46	82	82
PID	76.04	78	71
WBC	97.28	99	94
Hepatitis	85.16	66	90
Dataset	**GDLR**		
	Accuracy	**Specificity**	**Sensitivity**
Cleveland	100	100	100
PID	77.64	79	75
WBC	97.89	99	98
Hepatitis	97.5	97	94

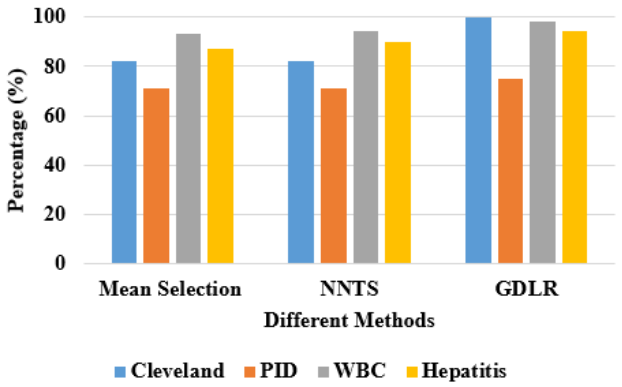

Figure.2 Performance of sensitivity

Compare to the traditional method the proposed GDLR method achieved better result in medical data classification. The major limitation of existing research technique includes both redundant and irrelevant features, hence it leads to less accuracy in the classification. In order to overcome this limitation, the proposed technique use the PCC and RF algorithm to select the important features for disease classification. So, the implemented technique reaches the maximum accuracy value compared to the other existing techniques. The graphical representation of sensitivity, specificity and accuracy is shown below.

The Fig. 2 shows the performance of sensitivity in proposed and existing medical data classification methods. The both MS and NNTS methods achieved 82% and 71% of sensitivity in Cleveland and PID dataset. Also, MS method achieved 93% and 87% of sensitivity and NNTS method achieved 71% and 94% of sensitivity in both WBC and hepatitis dataset. The

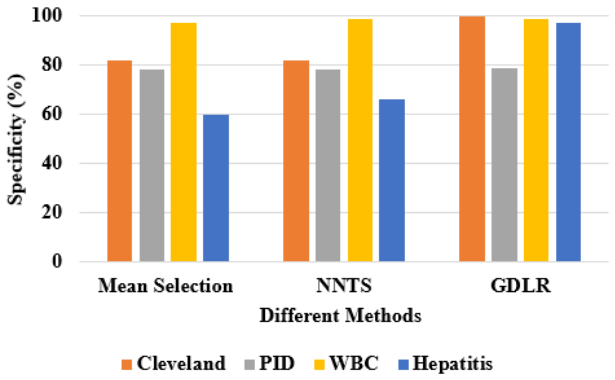

Figure.3 Performance of specificity

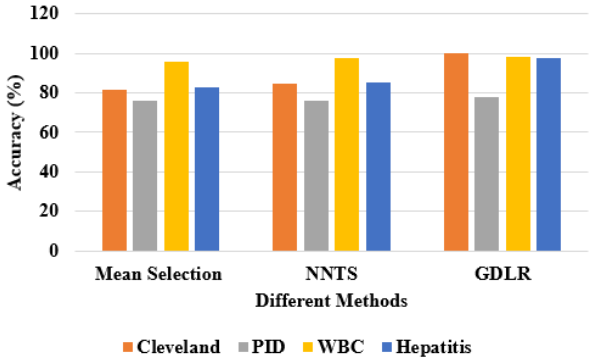

Figure.4 Performance of accuracy

GDLR classifier achieved 100%, 75%, 98%, and 94% of sensitivity with respect to Cleveland, PID, WBC and Hepatitis. The feature importance calculated by the RF is not efficient and the feature is applied to the GDLR technique to further analyze the feature importance and classify the data.

The Fig. 3 represents the performance of the specificity in the medical data classification system. The traditional MS and NNTS method achieved 82% and 78% of specificity with respect to Cleveland and PID datasets. The MS method achieved 97% and 60% similarly, NNTS method achieved 99% and 66% of specificity with respect to WBC and Hepatitis datasets. Moreover, GDLR classifier used different medical datasets such as Cleveland, PID, WBC, Hepatitis and it achieved 100%, 79%, 99% and 97% of specificity.

The Fig. 4 shows the performance of prediction accuracy in existing and proposed GDLR classifier. An existing MS algorithm achieved maximum classification accuracy as 95.99% in terms of WBC dataset and minimum accuracy as 81.75% in Cleveland dataset. The NNTS method achieved maximum 97.28% of accuracy in WBC dataset and minimum 76.04% accuracy in PID dataset. Finally, the GDLR classifier achieved maximum 100% of accuracy in Cleveland dataset and 97.89% of

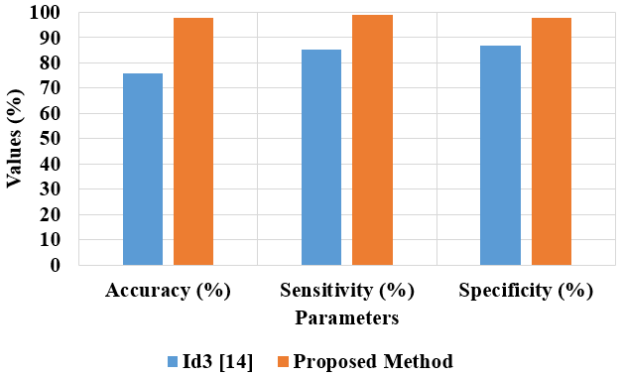

Figure. 5 Comparison of the proposed and Id3 [14] in WBC dataset

accuracy in WBC dataset. Compared to the existing methods the proposed GDLR classifier shows better results. Generally, the GDLR classifier performed faster than other classifiers. In addition, PCC and RF algorithm only select highly correlated features then forwarded to the GDLR classifier as a result, prediction accuracy improved.

The proposed method and the Id3 [14] method is evaluated in the WBC dataset and the parameter are measured. The accuracy, sensitivity and Specificity of the Id3 [14] and proposed method are shown in the Fig. (4). This shows that the proposed method has the higher performance compared to the Id3 technique due to the proposed method avoids overfitting problem in the data classification. The proposed method is interpretable compared to the existing method. Therefore, the proposed method has the higher performance compared to existing Id3 [14] method. The proposed method has the accuracy of 97.89% and the existing Id3 [14] has the accuracy of 75.83%.

4.3 Quantitative analysis

In this section, the Table 2 represents the comparative study of existing and the proposed GDLR based medical disease classification. The Cleveland datasets are used to measure the performance of the proposed GDLR and compared with other existing method [18, 19] in same dataset. The proposed GDLR method has the higher performance due to feature are selected by random forest are further analyzed by the GDLR to measure the feature weight value and classify the medical data based on the feature weight values.

The proposed GDLR method has the advantages of high interpretable and avoid overfitting. The GDLR method easily conduct the regularization and provide well-calibrated output. Hence, the GDLR classifier achieved better results compared to other

Table 2. Comparative study

Author Name	Methodology	UCI Medical Datasets	Parameters		
			Accuracy	Sensitivity	Specificity
Shaha, [18]	PPCA	Cleveland	85.82	80.43	88.42
S. Bashir, [19]	Hierarchical Majority Voting (HMV)	Cleveland	84.49	83.82	88.41
Proposed Work	GDLR	Cleveland	100	100	100

existing methods. The GDLR method achieved maximum 100% of accuracy with respect to Cleveland dataset and minimum 77.64% of accuracy in PID dataset.

The computation time of the proposed method in PID dataset is achieved as 154 ms, while existing method FOA-SVM method achieved as 170 ms. The proposed GDLR has the memory usage of 2,032 bytes and existing IID3 method [14] has the memory usage of 7,060 bytes.

5. Conclusion

The medical data analysis and classification are challenging issue in today's research because volume of medical data increasing rapidly. Selection of a suitable set of features makes it possible to classify an enormous quantity of data quickly and efficiently. In this research work, GDLR algorithm is proposed for medical data classification. The PCC method is employed to calculate the feature correlation and RF algorithm is employed for ranking the features. After that, GDLR classifier classifies the medical data as normal data and abnormal data. The GDLR method is applied in different UCI datasets such as Cleveland, WBC, PID and Hepatitis. The classification performance is estimated with the help of evaluation metrics such as sensitivity, specificity and accuracy. The proposed GLDR method has achieved accuracy of 97.89 in WBC dataset, while existing method NNTS method has the accuracy of 97.28 in same dataset. In future, research work can be extended to improve the classification performance of larger medical dataset using an improvised artificial intelligence approach like deep learning methodologies.

References

[1] N. S. Nithya and K. Duraiswamy, "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface", *Sadhana*, Vol.39, pp.39-52, 2014.

[2] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification", *Expert Systems with Applications*, Vol.41, No.5, pp. 2239-2249, 2014.

[3] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Medical data classification using interval type-2 fuzzy logic system and wavelets", *Applied Soft Computing*, Vol.30, pp.812-822, 2015.

[4] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification", *IEEE Journal of Biomedical and Health Informatics*, Vol.20, No.2, pp.655-668, 2016.

[5] A. Alsayat and H. El-Sayed, "Efficient genetic K-means clustering for health care knowledge discovery", In: *Proc. of IEEE International Conf. on Software Engineering Research, Management and Applications*, pp.45-52, 2016.

[6] J. M. Tomczak and M. ZięBa, "Probabilistic combination of classification rules and its application to medical diagnosis", *Machine Learning*, Vol.101, No.1-3, pp.105-135, 2015.

[7] T. Elguebaly and N. Bouguila, "A hierarchical nonparametric Bayesian approach for medical images and gene expressions classification", *Soft Computing*, Vol.19, No.1, pp.189-204, 2015.

[8] K. C. Lin and Y. H. Hsieh, "Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms", *Journal of Medical Systems*, Vol.39, No.10, pp.119, 2015.

[9] H. M. Harb and A. S. Desuky, "Feature selection on classification of medical datasets based on particle swarm optimization", *International Journal of Computer Applications*, Vol.104, pp.14-17, 2014.

[10] U. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records", *Computers & Electrical Engineering*, Vol.66, pp.487-504, 2018.

[11] L. Peng, H. Zhang, H. Zhang, and B. Yang, "A fast feature weighting algorithm of data

gravitation classification", *Information Sciences*, Vol.375, pp.54-78, 2017.
[12] M. Seera, C. P. Lim, S. C. Tan, and C. K. Loo, "A hybrid FAM–CART model and its application to medical data classification", *Neural Computing and Applications*, Vol.26, No.8, pp.1799-1811, 2015.
[13] Y. Xu, "Maximum margin of twin spheres support vector machine for imbalanced data classification", *IEEE Transactions on Cybernetics*, Vol.47, No.6, pp.1540-1550, 2017.
[14] S. Yang, J. Z. Guo, and J. W. Jin, "An improved Id3 algorithm for medical data classification", *Computers & Electrical Engineering*, Vol.65, pp.474-487, 2018.
[15] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets", *Springer Plus*, Vol.5, No.1, pp.1304, 2016.
[16] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, "Evolving support vector machines using fruit fly optimization for medical data classification", *Knowledge-Based Systems*, Vol.96, pp.61-75, 2016.
[17] P. Jaganathan, and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification", *Computers in Biology and Medicine*, Vol.43, No.12, pp.2222-2229, 2013.
[18] S.M.S. Shah, S. Batool, I. Khan, M.U. Ashraf, S.H. Abbas, and S.A. Hussain, "Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis", *Physica A: Statistical Mechanics and its Applications*, Vol.482, pp.796-807, 2017.
[19] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HMV: a medical decision support framework using multi-layer classifiers for disease prediction", *Journal of Computational Science*, Vol.13, pp.10-25, 2016.
[20] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.
[21] D. Dua and C. Graff, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]", Irvine, CA: University of California, School of Information and Computer Science, 2019.
[22] Y. Hayashi and S. Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset", *Informatics in Medicine Unlocked*, Vol.2, pp.92-104, 2016.
[23] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets", *Optimization Methods and Software*, Vol.1, No.1, pp.23-34, 1992.
[24] H. F. Eid, A. E. Hassanien, T. H. Kim, and S. Banerjee, "Linear correlation-based feature selection for network intrusion detection model", In: *Proc. of International Conf. on Advances in Security of Information and Communication Networks*, pp.240-248, 2013.