Refining Performance Metrics of Diabetes Prediction Using Enhanced Catboost Classifier in Big Data Analytics

G.Geo Jenefer [1*], Dr.A.J Deepa[2]
[1]Research Scholar, Anna University, Chennai, TamilNadu, India
[2]Professor, Ponjesly College of Engineering, Nagercoil, TamilNadu, India
E-mail:[1]geo.jenefer@gmail.com,[2]ajdeepajames@gmail.com

Abstract

Currently, diabetes plays a great challenge among human being. Diabetes is a disease that prevents our body from the proper usage of nutrients from food. It affects people of all ages. It is necessary to predict such kind of diseases at an earlier stage, treat and cure them. This can be done using Big Data Analytics. It also helps the medical practitioners to predict and take a decision on treating the disease. Many researchers have used different machine learning classifiers for early prediction of disease and differentiated them by their performance metrics. This paper proposes the Enhanced CatBoost classifier where Linear Discriminant Analysis (LDA) is used along with CatBoost Classifier for dimensionality reduction and to improve the performance. The objective of this work is to predict diabetes earlier in public health. The health data is taken from the PIMA diabetes dataset, and the results are compared. It was proven that the Enhanced CatBoost classifier produces the best accuracy of 96% with precision of 99%.

Keywords: *Big Data Analytics, CatBoost Classifier, Linear Discriminant Analysis, Machine Learning.*

I.Introduction:

1.1 Big Data:

Issues in managing huge data from hospitals can only be solved using Big Data. Medical practitioners use this to predict the disease and make decisions. It is a most challenging task to manage a huge amount of data [1].
It has five key features:
Volume – specifies the amount of data generated. It was predicted that more than 40 Zettabytes of data may get generated by 2020.
Source – specifies from where data is extracted. It increases as volume increases.
Velocity – specifies the rate of collected data used for analysis[1].
Variety – specifies the various types of organized and unorganized data from different sources[1].
Veracity – specifies the predicted values, data quality, reliability etc. [15].
Big Data Analytics: Big Data Analytics (BDA) is a complicated process that helps organizations to make business decisions.It includes components like predictive models, statistical algorithms etc. It analyses the data which are retrieved from warehouses and helps the business to decide their future outcomes. Hadoop Distributed File System in BDA helps to read and load faster. BDA applies machine learning algorithms for early prediction of diseases. It plays a best role in predicting the future health issues from the history of patients and provides better outcomes. The history of patients is used as the input. It trains the input dataset and evaluates using machine learning classifiers and detects the disease of patients earlier. It plays a key role in health care department.[3]

Various kinds of Big Data Analytics are:

1. Descriptive Analytics - It outlines the past data and converts it to human understandable form.
2. Diagnostic Analytics - It diagnoses the in-depth perception into a problem.
3. Predictive Analytics - It focus into the ancient and current data to make predictions of the upcoming.
4. Prescriptive Analytics - It prescribes solution to certain issues.

Big Data Analytics in Health Care : BDA in healthcare analyses large dataset which contains thousands of patient data and predicts the disease using data mining techniques. It provides instruments for management, analysis of large, diverse, structured and unstructured data [15].

Challenges : Major challenge in BDA is to make decisions based on predictive analysis [3]. Some other challenges faced by BDA in health care are: Obtaining High Throughput, data security, Sharing patient data with outside world etc. At times, choosing the best tools and platform make confusion while using Big Data Analytics.

Advantages: BDA helps to detect the disease earlier which provides a lot life-saving outcomes [15]. Medical Practitioners can take their own decisions. Also health care sector gets profit by BDA [3]. BDA leads to higher revenues and smarter business movements[3].

1.2 Diabetes

If the body is unable to produce its own insulin, there may be an increase in the blood sugar level. Such kind of disorder is known as Diabetes Mellitus. It affects the nervous system, causes more risk of stroke, kidney disease, heart attack and vision loss. Early prediction of this disorder helps to control blood sugar level and maintain health.

It was predicted that, by 2040, the count of diabetic patients in the world will be 642 million[7].

There are three types of diabetes given as follows:

Type 1: This disease affects the younger aged and people less than 30 years old. It is also called as Insulin dependent[7][8]. This damages the pancreas and so it stops the secretion of insulin[11]. Frequent urination, increased thirst, high blood glucose level etc., are some of the symptoms of Type 1 diabetes[7].

Type 2: It mostly affects the middle aged and the elderly people. It is also called as non-insulin dependent[7][8].This type of diabetes is caused due to the eating habits, over-weight, family history etc.[11]. Obesity, hypertension, arteriosclerosis, dyslipidemia are some of the symptoms of Type 2 diabetes[7].

Type 3: It is otherwise called as Gestational diabetes[9]. It increases the blood sugar level in pregnant ladies[11]. Increased blood sugar level causes heart and kidney diseases [10].

II.Research Studies:

Venkatesh et al., implemented Bigdata Predictive Analytics along with Naïve Bayes Technique (BPA-NB) for predicting disease. For handling huge dataset, Naïve Bayes was combined with Big Data Analytics. Dataset was taken from UCI repository and prediction was made on test data of the dataset. It was proven that it provided 97.12% accuracy. Hadoop spark was used as the tool.

Wang et al., proposed a new Convolutional Neural Network based Multimodal disease risk prediction algorithm for the prediction of chronic disease. Cerebral Infraction was considered as the chronic disease. Real time dataset was taken from Central China 2013-2015 for implementation. While execution, theproposed algorithm produced 94.8% accuracy.

Zou et al., used Decision Tree, Random Forest and Neural Network to predict Diabetes Mellitus. The dataset was taken from the hospital physical examination data in Luzhou, China. For dimensionality reduction, PCA along withminimum redundancy maximum relevance is used and it was shown that the prediction using RF obtained 80% accuracy.
Mujumdar et al., applied various machine learning algorithms on the dataset and it was proved that Logistic Regression provides 96% accuracy which was better than others. After applying pipelining to all the algorithms, Adaboost produced 98.8% accuracy. This work proved that accuracy and precision was improved after including pipelining.
Lai et al., implemented various machine learning models for the dataset taken from Canadian Population. They used Random Forest, Logistic Regression and Decision Tree algorithms for classification. Finally, it was proven that Logistic Regression provided better results than the Random Forest and the Decision Tree algorithms.
Sneha et al., applied Decision Tree, Naïve Bayes and Support Vector Machine algorithms for their classification. The specificity of DT was 98.2% and NB was 98%. In the proposed system, the accuracy was improved by selecting the optimal features of the dataset.
Nibareke et al., compared Linear Regression, Naïve Bayes and Decision Tree to predict diabetes. They implemented several Big Data tools. From these algorithms it was proven that DT performed better comparing to others with zero error.
Dinh et al., compared some machine learning algorithms like Linear Regression, Support Vector Machine, Random Forest and Gradient Boosting algorithm. They predicted cardiovascular, prediabetes and diabetes for datasets obtained from National Health and Nutrition Examination Survey. From these algorithms, for diabetes, the XGBoost with laboratory data achieved 95.7% AUROC and for prediabetes, AUROC was 73.7%.
Patil et al., compared various machine learning algorithms in PIMA diabetes dataset and proven that Gradient Boosting performed better with 79% accuracy, 77% precision, 74% recall, 75% f1 score and 75% ROC.
Al-Sarem et al., used various collaborative methods to find the most important feature for predicting parkinson's disease. CatBoostreached a highest accuracy of 86.25% with 23 features. RF obtained best accuracy with 31 features.
Ibrahim et al., used various machine learning algorithms and it was proven that CatBoost classifier achieved a highest score with 95% f-measure, 82% AUC and 91% precision. This algorithm is applied on loan approval and staff promotion.
Ghori et al., used collaborative methods, ANN and classifiers. The dataset was taken from Pakistan power supply company to detect and prevent Non-Technical Loss in power industry. It was found that the ensemble method scoredbetter for f-measure, ANN scored better for recall and CatBoostscored better as a classifier.
The above case studies were revised to use ML algorithms in a Predictive Analysis of Diabetic Patients.
The following table (Table 1) shows the usage of various machine learning algorithm in different dataset and their performance metrices are compared.

Table 1: Comparison of Performance Metrics of various Classification Algorithm from Existing Research Works.

Reference	Classification Algorithm	Disease	Performance Metrics	Dataset
Venkatesh et al., [16]	NB	Diabetes	Accuracy : 97.12%	UCI repository
Wang et al., [17]	CNN	Cerebral Infraction	Accuracy : 94.8%	Real time data from Central China 2013-2015

Zou et al.,[7]	RF with PCA	Diabetes Mellitus	Accuracy : 80%	Hospital Physical Examination Data in Luzhou, China
Mujumdar et al.,[8]	AB with pipelining	Diabetes Mellitus	Accuracy : 98.8%	-
Lai et al.,[9]	LR compared with RF,DT	Diabetes Mellitus	AUROC : 84% Sensitivity : 73.4%	Canadian Population
Sneha et al.,[11]	DT compared with SVM and NB	Diabetes Mellitus	Specificity : 98.2%	UCI Repository
Nibareke et al.,[18]	DT is compared with LR and NB	Diabetes Mellitus	Best accuracy with 0 error	-
Dinh et al.,[19]	XGBoost	Cardiovascular, Prediabetes, Diabetes	AUROC : 83.1% 73.7% 95.7%	National Health and Nutrition Examination Survey
Patil et al.,[21]	GB	Diabetes	Accuracy: 79% Precision:77% Recall: 74% f1 score: 75% ROC :75%	PIMA diabetes dataset.

The following table shows the usage of Catboost classifier in different dataset and their performance metrices are compared.

Table 2: Comparison of Performance Metrices of Catboost Classifier used in Different datasets from Existing Research Work with the Proposed Work.

Reference	**Performance Metrices**	**Dataset – Area**
Al-Sarem et al., [22]	Accuracy : 86.25%	UCI Repository – Parkinson's disease dataset
Rahman et al., [23]	Accuracy : 87.3%	Publicly available dataset - Waist-mounted smartphone
Ghori et al., [24]	Precision : 98% Recall : 99% F-measure : 98%	Power supply company in Pakistan – Non-Technical Loss in Power industry
Mamprin et al., [25]	Accuracy : 90% AUC_ROC : 83% F1_score : 45%	Catharina Hospital dataset – Heart disease
Ibrahim et al., [26]	AUC_ROC : 82% Precision : 91% F1-score : 95%	US Govt. - mortgage approvals.

Proposed work (ECB)	Precision : 99% Recall : 90% Accuracy : 96% F1_score : 94% Hamming Loss : 3% ROC_AUC : 94%	PIMA Diabetes Dataset

From the above comparisons it was proven thatexisting CB obtained accuracy of 86.25%, 87.3% and 90% in UCI Repository – Parkinson's disease dataset, Publicly available dataset regarding Waist-mounted smartphoneand Catharina Hospital dataset – Heart diseaserespectively. Also, it obtained precision of 98% and 91% in Power supply company in Pakistan regarding Non-Technical Loss in Power industry and US Govt. - mortgage approvals respectively.

The proposed ECB system provides 96% accuracy and 99% precision in PIMA Diabetes Dataset.

III. Proposed System

The proposed work examined the PIMA diabetes dataset with Enhanced Catboost Classifier (ECB).The Catboost Classifier is improved by including feature Scaling and LDA dimensionality reduction technique.

3.1 Proposed ECB system

The following schematic diagram(Figure 1) represents the architecture of the ECB system.

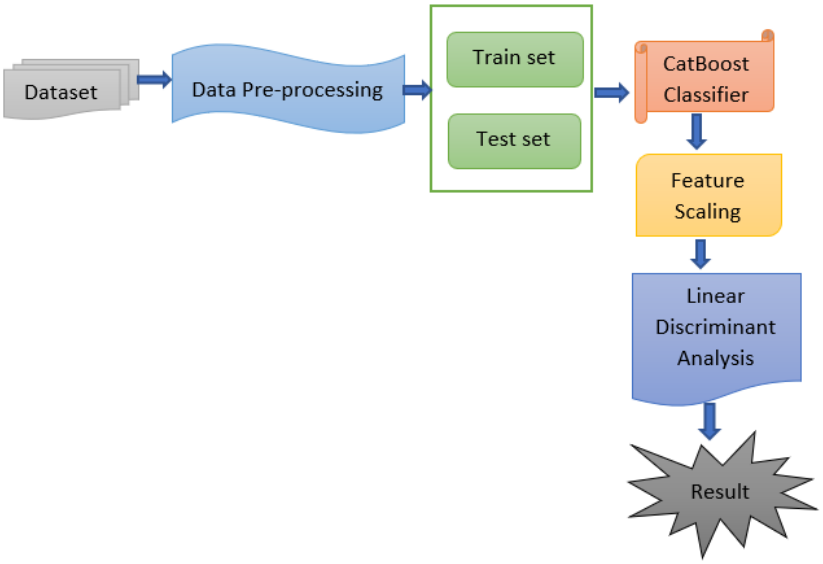

Figure 1: Proposed ECB System

Steps included in the ECB system

1. The input dataset is allowed for pre-processing. After preprocessing the data is separated as training and testing data.
2. By using ECB algorithm, the data is evaluated.
3. Then the split data is evaluated using ECB classification algorithms.
4. After evaluation, feature scaling is applied for normalizing the variables.
5. Finally, Linear Discriminant Analysis(LDA) is applied in the dataset for decreasing the dimensionality of the data.

Data Pre-processing

Data pre-processing is the first step done after getting the input data from dataset. It is done to replace the missing attributes and to normalize the values using normalization techniques. It is a datamining technique which converts the raw data into an understandable input format. Various techniques included in data pre-processing are:

1. Data Cleaning – It is the technique of removing the irrelevant or incomplete data.
2. Data Integration – It is the technique of combining the different sources into single form.
3. Data Transformation – It is the technique of converting the data from one format to another.
4. Data Reduction–It is the technique of reducing the original data.

In this work, the input data from the Pima Indian diabetes dataset is allowed for cleaning which discards outliers with mean value and is split into training set and test set. Data Pre-processing includes the following steps:

1. Reading input dataset as csv file to the data frame
2. Cleaning the input data by discarding maximum value for skin thickness
3. Replacing zero with mean of the variable
4. Splitting the dataset into features and labels. Features include all variables except outcome. Labels include only the outcome variable.
5. Splitting the features and labels into the train set and test set. A total of 767 samples in the dataset is split into 613 training set and 154 test set.
6. Write out all data

Classification Algorithm

CB Classifier

It is one of the machine learning algorithms available as an open source.Catboost performs one-hot encoding to convert categorical data into numerical data. Instead of values, CB takes the indices of categorical data. CB depends on gradient boosting algorithm.

The conversion of categorical data into numerical data takes place according to the following three steps:

1. Randomly permutate the input data number of times.
2. Now, the label values are converted from categorical value to integer value.
3. Finally, the categorical features are converted into numerical values.

Feature Scaling

It is used to normalize the range of independent variables. It lies between the maximum and the minimum value (0-1). For each feature, the maximum absolute value is scaled to unit size.

The general formula for min-max of [0,1] is given as,

$$f1 = f - \frac{\min(f)}{\max(f)} - \min(f) \qquad (1)$$

Where f is the original value and f1is the normalized value.

To rescale a range between an arbitrary set of values (x,y), the following min-max normalization is used.

$$f1 = x + (f - \min(f))(y - x)/\max(f) - \min(f) \qquad (2)$$

Where x and y are the min and max values.

Mean Normalization can be calculated as

$$f1 = f - \frac{average(f)}{\max(f)} - \min(f) \qquad (3)$$

Linear Discriminant Analysis

It is a dimensionality reduction technique used along with classification algorithm to improve its performance. Mostly, LDA is used in predictive analysis and image

recognition. It reduces the dimensionality of the dataset by removing the redundant features.

Pseudocode:

1. *Get the input data from the dataset.*
2. *Calculate the d-dimensional mean vector for each class in the dataset. Mean can be calculated as,*

$$m = \frac{1}{N}\sum_{x=1}^{k} mxy$$

Where N denotes the sample data.

3. *Calculate the scatter matrices for in-between-class and within-class.*
   3.1 Scatter Matrix for in-between class can be calculated as,

$$Sb = \sum_{x=1}^{k} Nx(mx - m)(mx - m)$$

3.2 Scatter Matrix for within-class can be calculated as,

$$Sw = \sum_{x=1}^{k} (Nx - 1)\sum x$$

4. *Find the eigen vectors(v1,v2,v3….vn) and eigen values(λ1,λ2,λ3,….λn) for each eigen vectors for the Scatter matrices.*
5. *Sort and rank the eigen vectors and choose thefirst largest v eigen vectors with corresponding values.*
6. *Construct dimensional matrix, W = v x d*
7. *Use the constructed matrix to transform the samples into the new subspace. This can be given as,*

$$Y = X\ x\ W$$

Where Y denotes n x v dimensional samples in new subspace, X denotes the n x d dimensional matrix.

IV. Performance Evaluation

Confusion Matrix: The performance of the classifier is described in a table called as Confusion Matrix. It is a performance measure for all machine learning classifiers. Each row and column in the Confusion Matrix table is called as actual class and predicted class respectively. The following table shows the terms that are provided in the confusion matrix.

Table 3: Terms used in confusion matrix

		Actual Values	
		1	0
Predicted Values	1	True Positive -T_rP_o (Predicted as diabetes and actually diabetes)	False Positive - F_aP_o (Predicted as diabetes but not actually diabetes)
	0	False Negative - F_aN_e (Predicted as non-diabetes and actually diabetes)	True Negative - T_rN_e (Predicted as non-diabetes and actually non-diabetes)

Precision / Recall TradeOff : Precision denotes the relationship between the patients actually having diabetes and predicted to have diabetes. Eq. 4 shows the precision tradeoff. In this work, the precision for ECB algorithm is 0.99.

$$Precision = TrPo/(TrPo + FaPo) \quad (4)$$

The recall denotes the relationship between patients who had diabetes diagnosed as having diabetes. Eq.5 shows the recall tradeoff.

$$Recall = \frac{TrPo}{(TrPo+FaNe)} \quad (5)$$

When the threshold value decreases, recall gets increases and precision may decrease. The measured recall tradeoff for ECB algorithm is 0.90.

Accuracy: Accuracy (AC) defines the number of precise predictions made by the model.

$$AC = \frac{(TrPo+TrNe)}{(TrPo+FaPo+TrNe+FaNe)} \quad (6)$$

After applying these classification algorithms, the accuracy of the diabetes dataset was found. The measured accuracy for ECB is 0.96.

F1 Score: It defines the harmonic mean of precision as well as recall. It is the weighted average of the precision as well as recall. The best value of F1 would be 1 and the worst value of F1 would be 0.

F1 Score can be calculated as,

$$F1 = \frac{2*(precision*recall)}{(precision+recall)} \quad (7)$$

The measured f1-score forECB algorithm is 0.94.

Hamming Loss: Hamming loss is the amount of incorrectly predicted values from the normal values. The value ranges from 0 to 1

The measured hamming loss forECB algorithm is .03. Hamming loss should be less for better classifier.

ROC_AUC: It indicates the Receiver or Operating Characteristic Curve. The curve value should be from 0 to 1. 0.5 is considered as the baseline and always the value must be more than 0.5. The measured ROC_AUC forECB algorithm is 0.94.

The Performance metrics mentioned above denotes that ECB results in highest accuracy with lowest Hamming Loss.

V. Experimental Results & Discussion

Dataset: The dataset of this work has been taken from the Pima Indian dataset which is at UCI ML Repository. It contains the details of diabetic as well as non-diabetic patients with eight attributes and one outcome. Totally 768 records are included in the dataset. The dataset is described in Table 4.

Table 4: PIMA Indian Diabetes dataset

Sl.No.	Attribute Name
1	Pregnancy(No. of times)
2	Level of Glucose
3	Blood Pressure(BP) Level (mm Hg)
4	Skin Fold Thickness (mm)
5	Insulin (mu U/ml)
6	Body Mass Index(BMI) (weight in kg, height in mm)
7	Diabetes Pedigree Function
8	Age (years)

9	Outcome (0 or 1)

The total 768 records are split as 614 train set samples and 154 test set samples.

In pre-processing step, max value for Skin Thickness attribute is removed and zero values are replaced in attributes such as Glucose, BloodPressure, Skin Thickness, Insulin, BMI and outcome. Now, the dataset is divided into features and target label. All the attributes except Outcome is considered as features and the Outcome is considered as the target label. Now, the features are divided as 613 training set and 154 test set.

The separated train and test set are allowed for CatBoost Classifier. Then it is allowed for data cleaning. The following (Table 5) shows the CB train set after cleaning. All performance metrices of CB after cleaning is described in the table and its graphical representation is shown in the figure(Figure 2).

Table 5: CB train set after cleaning

Sl.No.	Performance Metrices	CB training set after Cleaning
1	Precision	0.975
2	Recall	0.895
3	Accuracy	0.954
4	F1_score	0.933
5	Hamming Loss	0.045
6	ROC_AUC	0.941

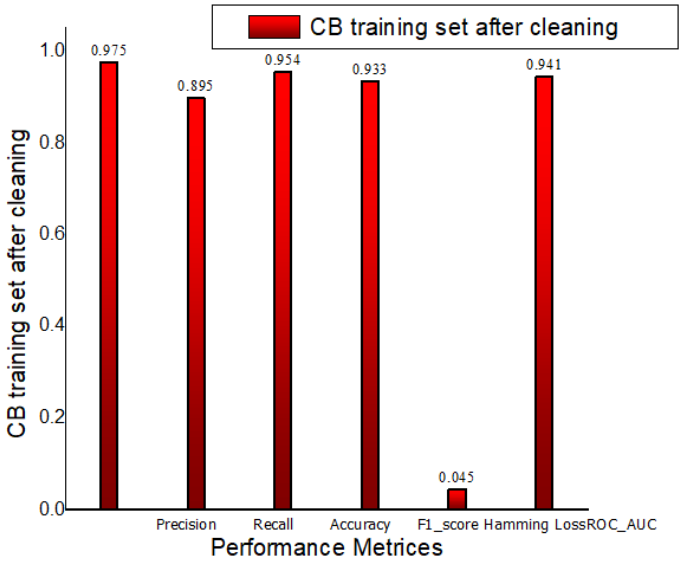

Figure 2: Graphical representation of CB train set after cleaning

The following table (Table 6) shows the performance metrices of CB test set after cleaning. From the below table it was noted that the performance metrices improved after cleaning the data. The graphical representation of this table is shown in the figure (Figure 3)

Table 6: CB test set after cleaning

Sl.No.	Performance Metrices	CB test set after Cleaning
1	Precision	0.711
2	Recall	0.680
3	Accuracy	0.818
4	F1_score	0.695
5	Hamming Loss	0.181
6	ROC_AUC	0.77

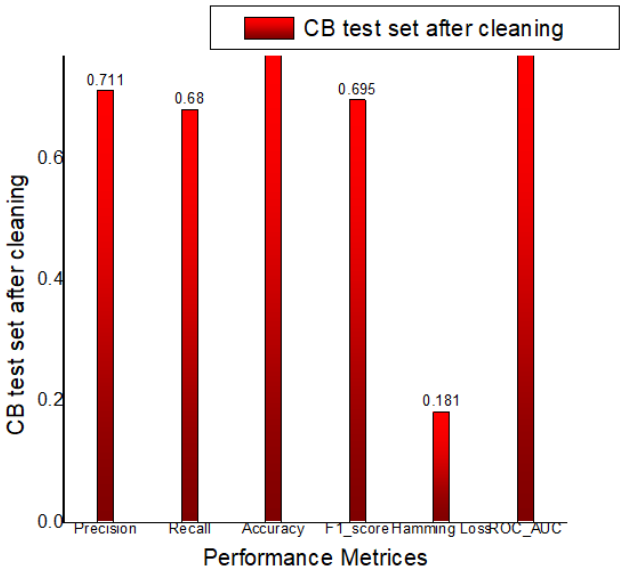

Figure 3: Graphical representation of CB test set after cleaning

Performance metrices of CB such as precision, recall, accuracy, F1_score, Hamming Loss and ROC_AUC produces greater values when including LDA dimensionality reduction. The following table (Table 7) show the performance metrices of CB train set after LDA. The graphical representation of this table is shown in the figure (Figure 4).

Table 7: CB train set after LDA

Sl.No.	Performance Metrices	CB train set afterLDA
1	Precision	0.689
2	Recall	0.869
3	Accuracy	0.844
4	F1_score	0.769
5	Hamming Loss	0.155
6	ROC_AUC	0.851

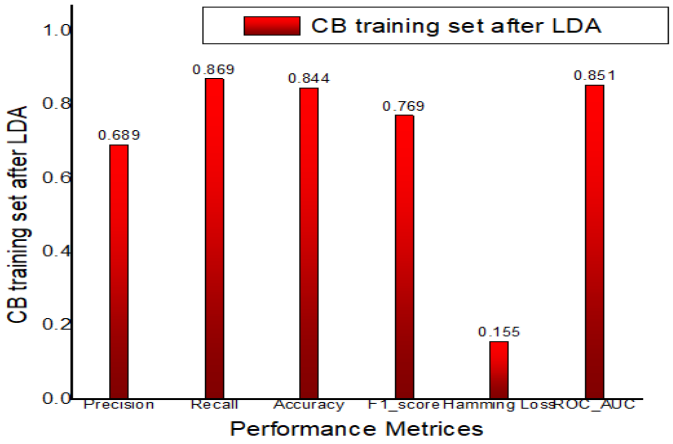

Figure 4: Graphical representation of CB train set after LDA

The following table (Table 8) show the performance metrices of CB test set after LDA. In accordance with the training set, the test set is evaluated and the results is described in the table. The Graphical representation of CB test set after LDA is shown in the figure (Figure 5)..

Table 8: CB test set after LDA

Sl.No.	Performance Metrices	CB test set after LDA
1	Precision	0.990
2	Recall	0.904
3	Accuracy	0.962
4	F1_score	0.945
5	Hamming Loss	0.037
6	ROC_AUC	0.949

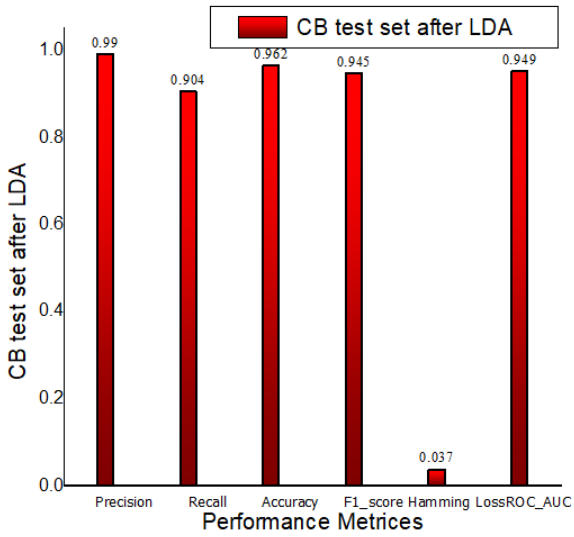

Figure 4: Graphical representation of CB test set after LDA

From the above table and graph it was proven that ECB (CB after LDA) performed better and produced best values.

Conclusion

Several classification algorithms were used for predicting diabetes in various research works. But they still didn't achieve better results in performance. In this ECB system, LDA dimensionality reduction technique is used along with classifier which reduces the data and improves the performance. This proposed system is used in the pima Indian diabetes dataset, and the performance is measured using various performance metrics like precision, recall, accuracy, hamming loss and ROC_AUC. Hence, it was proven that ECB results in better prediction of diabetes with more accuracy and less hamming loss. The performance of the ECB system increases by increase in all the metrics and decrease in hamming loss.

Future Enhancement:

In the future, the image dataset can be used to predict diabetes disease. Also, this ECB system can be used in prediction of various chronic diseases like lung disease, liver disease and so on. It can also be used in the prediction of COVID-19.

Acknowledgment

This research was supported by Dr. Deepa A J Professor in the Department of CSE at Ponjesly College of Engineering, Nagercoil. We are thankful to my brother G.Geo Niju Shanth [Lead Consultant] who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper. We are also grateful to our friends for assistance with [Big Data Analytics], and who moderated this paper improved the manuscript significantly.

Funding: I did not apply any funds for our research

Conflict of Interest

No conflict of interest.

References

1. *Dash, Sabyasachi, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. "Big data in healthcare: management, analysis and future prospects." Journal of Big Data 6, no. 1 (2019): 1-25.*
2. *Onyemachi, NkemakolamChinenye, and Ogwueleka Francisca Nonyelum. "Big Data Analytics in Healthcare: A Review." In 2019 15th International Conference on Electronics, Computer and Computation (ICECCO), pp. 1-5. IEEE, 2019.*
3. *Reddy, A. Rishika, and P. Suresh Kumar. "Predictive big data analytics in healthcare." In 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), pp. 623-626. IEEE, 2016.*
4. *Hassan, Ch Anwar Ul, Muhammad Sufyan Khan, and Munam Ali Shah. "Comparison of Machine Learning Algorithms in Data classification." In 2018 24th International Conference on Automation and Computing (ICAC), pp. 1-6. IEEE, 2018.*
5. *Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19, no. 1 (2019): 1-16.*
6. *Grampurohit, Sneha, and Chetan Sagarnal. "Disease Prediction using Machine Learning Algorithms." In 2020 International Conference for Emerging Technology (INCET), pp. 1-7. IEEE, 2020.*

7. *Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.*
8. *Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.*
9. *Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. "Predictive models for diabetes mellitus using machine learning techniques." BMC endocrine disorders 19, no. 1 (2019): 1-9.*
10. *Sarwar, Muhammad Azeem, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. "Prediction of diabetes using machine learning Algorithms in healthcare." In 2018 24th International Conference on Automation and Computing (ICAC), pp. 1-6. IEEE, 2018.*
11. *Sneha, N., and TarunGangil. "Analysis of diabetes mellitus for early prediction using optimal features selection." Journal of Big data 6, no. 1 (2019): 1-19.*
12. *G. Geo Jenefer, Dr.A.J. Deepa "A Cognitive Survey on Big Data Analytics in Predicting Chronic Diseases" Journal of Computational Information Systems Volume 14 - Issue 6 December 2018.*
13. *Deepa, A. J., and V. Kavitha. "Efficient intrusion detection system using random tree." International Journal of Enterprise Network Management 6, no. 4 (2015): 275-285.*
14. *Supriya, M., and A. J. Deepa. "A novel approach for breast cancer prediction using optimized ANN classifier based on big data environment." Health care management science (2019): 1-13.*
15. *Ristevski, Blagoj, and Ming Chen. "Big data analytics in medicine and healthcare." Journal of integrative bioinformatics 15, no. 3 (2018).*
16. *Venkatesh, R., C. Balasubramanian, and M. Kaliappan. "Development of big data predictive analytics model for disease prediction using machine learning technique." Journal of medical systems 43, no. 8 (2019): 1-8.*
17. *WANG, LU, and LIN WANG. "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities."*
18. *Nibareke, Thérence, and Jalal Laassiri. "Using Big Data-machine learning models for diabetes prediction and flight delays analytics." Journal of Big Data 7, no. 1 (2020): 1-18.*
19. *Dinh, An, Stacey Miertschin, Amber Young, and Somya D. Mohanty. "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning." BMC medical informatics and decision making 19, no. 1 (2019): 1-15.*
20. *Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50 (2015): 203-208.*
21. *Patil, Ratna, and SharavariTamane. "A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes." International Journal of Electrical and Computer Engineering 8, no. 5 (2018): 3966.*
22. *Al-Sarem, Mohammed, Faisal Saeed, WadiiBoulila, Abdel Hamid Emara, Muhannad Al-Mohaimeed, and Mohammed Errais. "Feature Selection and Classification Using CatBoost Method for Improving the Performance of Predicting Parkinson's Disease." In Advances on Smart and Soft Computing, pp. 189-199. Springer, Singapore, 2020.*
23. *Rahman, Saifur, Muhammad Irfan, Mohsin Raza, Khawaja Moyeezullah Ghori, Shumayla Yaqoob, and Muhammad Awais. "Performance analysis of boosting classifiers in recognizing activities of daily living." International journal of environmental research and public health 17, no. 3 (2020): 1082.*

24. *Ghori, Khawaja Moyeezullah, Rabeeh Ayaz Abbasi, Muhammad Awais, Muhammad Imran, Ata Ullah, and Laszlo Szathmary. "Performance analysis of different types of machine learning classifiers for non-technical loss detection." IEEE Access 8 (2019): 16033-16048.*
25. *Mamprin, Marco, S. Zinger, P. H. N. de With, J. M. Zelis, and P. A. L. Tonino. "Gradient boosting on decision trees for mortality prediction in transcatheter aortic valve implantation." In Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology, pp. 325-329. 2020.*
26. *Ibrahim, Abdullahi, Muhammed M. Muhammed, Samuel O. Sowole, Ridwan Raheem, and Rabiat O. Abdulaziz. "Performance of CatBoost classifier and other machine learning methods."*