LGBM Classifier based Technique for Predicting Type-2 Diabetes

B. Shamreen Ahamed,
Department of Computer Science and Engineering,
Research Scholar,
SRM Institute of Science and Technology,
Vadapalani Campus, Chennai
sb3319@srmist.edu.in

Dr. Meenakshi Sumeet Arya
Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
SRM Institute of Science and Technology,
Vadapalani Campus, Chennai
meenaksa1@srmist.edu.in

Abstract

In today's world, Diabetes Mellitus is a disease, that is considered to be an extensive noncommunicable disease which has a great effect our day to day living. In the 21st century, changes in natural life style and labor culture are some of the main reasons for India to have 62 million diabetic cases as of today. Analytical Computational Techniques can be applied on clinical immense data, the enormous quantity of data produced in the healthcare schemes, there is a option to form medicinal intelligence which will initiative medical forecast and predicting in future. By advancing medical intelligence and with the help of development model, prediction and detection of diabetes disease can be done. With the increase in complexity to the problems, the accuracy percentage also varies. LGBM - Light Gradient Boosting Algorithm is one such algorithm that can be used as it depends on decision tree algorithms and it can be used in predicting the accuracy to attain the desired results. With the existing PIMA Indian Dataset the accuracy is calculated as 95.20% using LGBM Algorithm . Therefore by using the LGBM classifiers, we can develop a data model for diabetes detection and prediction.

Keywords: *Diabetes Mellitus, LGBM, PIMA Indian Dataset.*

1. INTRODUCTION

In today's digital world, there are a number of chronic auto immune diseases that affect the everyday growth of an individual. Diabetes is one such disease that has diverse effects in the human health and from the various advanced technologies that have emerged in the recent medical field, the death toll caused by the diabetes disease can be reduced[1]. Big Data Analytics is used as a tool to design and analyze the data, manage and precisely obtain the information that is needed from huge amount of data sets that consist of data that is similar to a particular patient within less amount of time. Some of the commonly used tools in software include advanced analytics strategy such as data mining, statistical analysis, predictive analytics and text analytics. In addition to this, the latest advanced analytical technologies can be used to transform the healthcare industry for better medical facilities at the appropriate time [2].

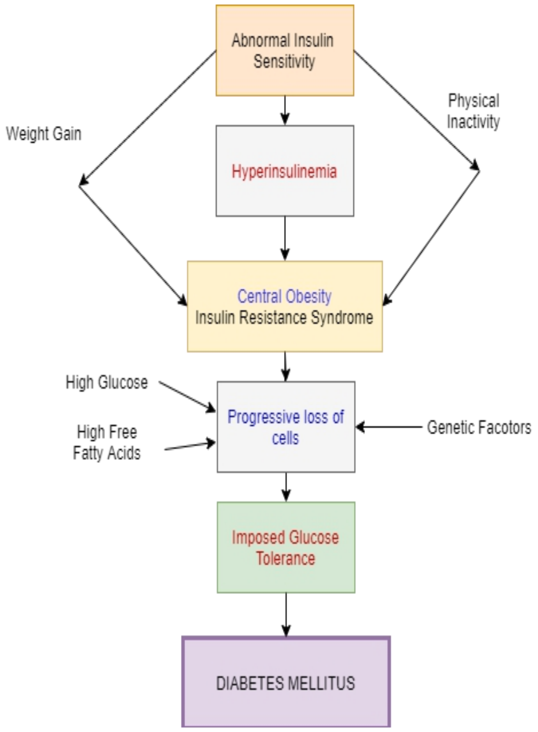


Fig 1.1 Flowchart of Diabetes Mellitus

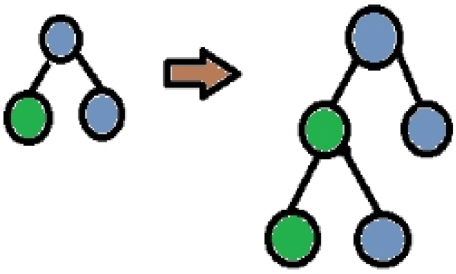
One major drawback that occurs when handing large amount of data is that there is no proper channeling of the data available in the environment. The hospital management system has different types of data such as numerical, statistical, raw, etc. This data needs to be categorized into different variants and categorized according to the medication given to the patients. The lack of such categorization has affected the result analysis to identify the causes and effects of the diseases. To overcome this problem a framework needs to be developed and enhanced [3].

As of today, the Diabetic Disease has affected many people and a majority of them still do not have a clear idea about the prevention to be taken and the risks involved in Diabetes [4].

Diabetes mellitus is a disease in which the maltose level in the blood cannot be maintained by the body. This is a type of a metabolic diseases in which the blood of a person has high maltose level and the insulin content produced does not have any response by the cells. This high blood maltose causes some early-symptoms such as frequent thirst, tiredness, itchy skin and headache. The main types of diabetes mellitus can be divided into two categories: namely Type- 1 diabetes and Type-2 diabetes. The Type- 1 Diabetes occurs in around 10% of people having diabetes [5]. This commonly affects children and adolescents. In this, the person is injected insulin periodically. During Type-2 diabetes, it mostly appears in aged people. In this case, the body does not make use of insulin appropriately[6]. Therefore the affected person has to use proper medications, diet, exercise and maintain their blood sugar levels at regular time intervals. The factors that disturb both Type-1 and Type-2 Diabetes is Genetic and environmental factors, but by having a healthy lifestyle choice, the Type-2 Diabetes can be avoided in many people. [7].

From this we infer that Diabetes is a disease that can be overcome if technology is used wisely and effectively. In today's technology, there are diverse kinds of appliance algorithms that can be used. These include classification algorithms or association algorithms [8]. Other algorithms such as Decision Trees(DT), Support Vector Machine(SVM), and Linear Regression(LR) can be used in predicting disease. However, no such data model exists for completely predicting, detecting and monitoring the disease. If one such model is developed the disease can be overcome in future[8].

This paper provides information regarding the same gives details about the LGBM Algorithm that can be used to determine and develop the data model. "The LGBM Algorithm depends on decision tree algorithms, it divides the tree leaf wise with the best fit though other boosting algorithms split the tree profundity wise or level wise as opposed to leaf-wise" [9].



Leaf wise Tree growth

In this, the leaf is considered as one type of data and its branches is considered as its sub groups. Therefore, when developing on a leaf in Well-lit GBM, the leaf-wise algorithm will be more accurate and brings about improved meticulousness that can be used to fetch proficient information by using any of the presently obtainable improving algorithms. Furthermore, it is tremendously swift, subsequently the term 'Light'[10].

The paper categorizes this as follows:

Section 1 gives a basic information about the Diabetes Disease, Section 2 provides information regarding the techniques used so far, Section 3 provides details about the related work done under diabetes disease, Section 4 focuses on the algorithm used and how Diabetes can be predicted and controlled using data analytics. In section 5, the results are produced. The conclusion and future work are discussed in Section 6.

2. RELATED WORK

In the health field, many ML algorithms are used for envisaging diseases. Diabetes is one such ailment that uses ML practices to foresee diabetes with the most accurate solutions.

Mercaldo et al.[14] cast-off 6 classifier algorithms. The classifier algorithms are JRip algorithm, J48 algorithm, BayesNet algorithm, Multilayer Perceptron algorithm, Hoeffding Tree algorithm and Random Forest algorithm. The study was intended for Pima Indian dataset[14].

The writers have taken two greatest algorithms certainty Greedy Stepwise and Best Initial, to verify the discriminatory qualities that aid in refining the concert classification. From plasma glucose concentration, BMI, age and diabetes pedigree function four characteristics were taken specifically. According to the authors, “a ten – fold - cross veradication is pragmatic to the dataset. The classifiers were allied and were thru liable on the value of the recall, accuracy and the F-Measure. The result showed the accuracy in the value that was equivalent to 0.757, F-measure equals to 0.759 and recollection equals to 0.762. From the Hoeffding Tree algorithm we can conclude that it formed the highest presentation linked to others”[14].

Sisodia et al. [15] used SVM algorithm, Decision Tree algorithm and Naive Bayes classifiers algorithm to predict the diabetes disease. The main objective was recognizing the classifier that has the highest accurateness. The dataset used is Pima Indian dataset. In the author’s opinion, “the 10-folds cross-validation partition was done. The performance was evaluated using the measures of the recall, accuracy, the precision and the F-measure. The Naive Bayes obtained highest accuracy, measuring 76.30%”[15].

Kandhasamy et al. [16] castoff multiple classifiers Random Forest, SVM, J48 and K- Nearest Neighbours (KNN). The classification was tested on UCI repository dataset. The outcomes founded on the various answers had to be calculated as to how specific, accurate and sensitive the classifiers were all related. The division was thru in a couple of diverse cases, with the dataset being pre-processed and not pre-processed by using five-fold cross edification. The novelists did not clarify the pre-processing phase practical on the dataset, It was believed that the J48 classifier algorithm, demonstrated highest accuracy rate of 73.82% without pre-processing, Random Forest demonstrated a highest accuracy rate of 100% with pre-processing enabled[16].

Tafa et al. [17] introduced a better improvised model of Naive Bayes and SVM for detecting diabetes disease. The data model functioned using a dataset that was obtained from 3 places in Kosovo state. The data collected contained 402 patients and 8 attributes where 20 percent of people had type-2 diabetes disease. Some of the attributes have not been investigated in the past. Some of them include the physical activity, regular diet and hereditary of diabetes. The authors The pre-processing status of the data was not given by the authors. For the validation test, “the dataset was split into 50% for each as the training and testing datasets. With the proposed combined algorithms an accuracy of 97.6% has been achieved. This value was compared with the performance of Naive Bayes and SVM achieving 94.52% and 95.52%”[17].

Yuvaraj et. al. [18] proposed a model for diabetes using 3 different ML algorithms including Decision Tree algorithm, Random Forest algorithm and the Naïve Bayes algorithm. The authors used, “a pre-processed Pima Indian Diabetes dataset (PID) was used. The authors did not mention how the data was pre-processed, however the Information Gain method used for feature selection to obtain the relevant features were used. 8 main attributes of the 13 were used. Also, the dataset was split into 70% and 30% for training and testing respectively. The results showed that the most accurate algorithm was random forest at the rate of 94%”[18].

Olaniyi et al. [19] used a Multilayer Feed-Forward Neural Network in their paper. In this, the “back-propagation algorithm" was utilised. The main aim was to predict diabetes more accurately. The Pima Indian Diabetes dataset was used . To achieve a numerical stability, the dataset was normalized before processing the classification. All dataset values were between zero and one which was achieved by dividing each sample attributes by their corresponding amplitude. After that, the “dataset was divided into 500 samples for a training set and 268 for the testing set. 82% was the highest accuracy” got[19].

Soltani et al. [20] used the “Probabilistic Neural Network (PNN) to predict diabetes disease”. The algorithm was applied to the “Pima Indian dataset”. The pre-processing technique was not apply by author. However, “the dataset is alienated into 10% for the setting customary and 90% for the training set. The projected technique attained exactness of 81.49% for testing and 89.56%, for taxing data”[20].

Rakshit et al. [21] used Pima Indian dataset for their study and predicted diabetes by obtaining Two-Class Neural Network.

According to the author, “the dataset was processed by normalizing all the sample attributes values using the standard deviation and mean of each attribute to obtain a numerical stability”. In accumulation, they found the pertinent structures using correlation. However, the authors didn’t mention these prejudiced geographies. The dataset was part into a training dataset and testing dataset in the ratio of “314:78” samples. The outcome of this model attained the maximum accuracy of 83.3%[21].

Three supervised learning algorithms were applied by Mamuda et al.[22]. They include Scaled Conjugate Gradient (SCG), Levenberg Marquardt (LM) and Bayesian Regulation (BR). This work is used by the Pima Indian dataset for assessing the performance. The “10-fold cross authentication was used to split the data into training data and testing data for validation study. The authors reported that Levenberg Marquardt (LM) obtained the best routine on the justification set based on the Mean Squared Error (MSE)” which equaled to 0.00025091[22].

Negi et al. [23] aimed to apply the concept of SVM to foresee diabetes. “The Diabetes 130-US and Pima Indian datasets were used in a combined form. The aim of this study was to legalize the unswerving outcomes as other researchers often used a solitary dataset. The dataset consists of 49 attributes and 102,538 samples 64,419 were positive models and 38,115 were negative samples. The dataset is pre-processed and normalized between 0 and 1 by swap the missing values and out of series data by zero. Various feature selection methods were tested before applying the SVM model. 4 attributes were selected by the F select script from LIBSVM package and the Wrapper and Ranker methods certain nine and twenty attributes, respectively. The authors used 10-fold cross validation technique for the validation process”. From the combined dataset, the diabetes detection might be extra reliable, with an accurateness of 72% to end[23].

John Martinsson et al.[24] presented a bottomless neural network prototypical that prophesied blood glucose levels simplifying an estimate level of vagueness in the prediction. By using the T1DM Dataset the glucose level was considered. The method is evaluated using the root-mean-square-error (RMSE) metric and Surveillance-error-grid (SEG) technique [24].

Jayanthi et al. [25] have ended a predictive examination using seven sorts of regression models as polynomial regression, logistic regression algorithm, linear regression algorithm, stepwise regression, ridge regression, Lasso regression algorithm and elastic net regression algorithm. The notion of existing "predictive models" and "clinical predictive models" is explicated in this paper. Though, in future the truth of the present system can be developed to a greater extent[25].

Guolin et al. [26] have future the LGBM Algorithm in their paper that has dual techniques : "Gradient Based one sided Sampling with Exclusive Feature Bundling that deals with large instances of data and performs theoretical analysis of it and produces results"[26].

3. BASIS OF LGBM

LGBM stances for Light Gradient Boosting Machine . The LGBM Algorithm uses two concepts. They are GBDT (Gradient Boosting Decision Tree) or GOSS (Gradient based one-sided sampling). These are the widely-used machine learning algorithm involved for prediction of study[27].

The GBDT is a boosting algorithm where, Boosting is a over-all collective process that replicates a sturdy classifier from a number of puny classifier. This is achieved by constructing a model framework from the working out data, then producing a 2nd model that goes to rectify the mistakes from the initial model framework. The efficacy, correctness and interoperability are its crucial factors. It is a advancing tree agenda that is very effectual and accessible[27] .

In recent years, "with the rise of big data, GBDT is facing new-fangled challenges, specially in the trade-off between proficiency and precision. Conventional implementations of GBDT scans all the data instances that is needed to obtain the required information. This will carry in a proportionality amid the number of cases and the number of structures involved"[27].

Another algorithm that can be involved through prediction of a model is GOSS. Gradient-based One-Side Sampling (GOSS) can also be used sideways the LGBM Algorithm concept. Although "there is no instinctive heft for data instance in GBDT, we notice that data instances with different gradients (The rate of rise or fall along the length of the road with respect to the horizontal is called grade or gradient) play varied roles in the computation of data exchange". Hence, when down sampling is done on the data instances, "large gradients must be used in order to hold the precision of information gain estimation" [28].

3.1 **Theoretical Analysis**

GBDT uses the concept of verdict trees inorder to cram a function is given as follows: "from the input space X s to the gradient space G [1]. A training set with instances {x1, · · · , xn} are assumed, where each xi is a vector with dimension s in space X . In each restatement of gradient boosting, the negative gradients of the loss function with respect to the output of the model are denoted as {g1, · · · , gn}"[28]. "The decision tree model divides each node at the most revealing feature (which gives rise to the largest evidence gain). In GBDT, the data improvement is measured by the variance after segregating", which can be explained as below[28].

*"**Y**=Base_tree(**X**)-lr*Tree1(**X**)-lr*Tree2(**X**)-lr*Tree3(**X**)"*

"Definition: Let O be the training dataset on a fixed node of the decision tree. The variance gain of dividing measure j at point d for this node is defined as

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\left(\sum_{\{x_i\in O:x_{ij}\le d\}} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{\{x_i\in O:x_{ij}>d\}} g_i\right)^2}{n^j_{r|O}(d)}\right)$$

where $n_O = \sum I[xi \in O]$, $n^j_{l|O}(d) = \sum I[xi \in O : xij \le d]$ and $n^j_{r|O}(d) = \sum I[xi \in O : xij > d]$. This is ended by using the concept of GBDT".

Gradient One-Sided Sampling (GOSS) uses every instance with larger gradient and performs random sampling on the instances with smaller gradients. The training dataset is given as O on a specific node of the decision tree algorithm. "The variance gain of dividing measure j at point d for this node is defined as

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i + \frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum_{x_i\in A_r} g_i + \frac{1-a}{b}\sum_{x_i\in B_r} g_i\right)^2}{n^j_r(d)}\right)$$

where A_l = {xi ∈ A : xij ≤ d}, A_r = {xi ∈ A : xij > d},B_l = {xi ∈ B : xij ≤ d}, B_r = {xi ∈ B : xij > d}, and the coefficient $\frac{1-a}{b}$ is used to normalize the sum of the gradients over B back to the size of A^c. This is done by using the concept of GOSS"[29].

Histogram based:

A histogram is used to abridge discrete or continuous data. In other words, it offers a visual interpretation of mathematical data by showing the sum of data points that fall within a quantified series of values (called "bins"). It is alike to a vertical bar graph[29].

Another way of implementing the algorithm is called the Histogram-based approach[29]. The details about this algorithm is given below:

Algorithm : Histogram-based[29]

```
"Input: I1: training dataset, dp: max depth
Input: m1: feature dimension
nodeSet1 ← {0}
rowSet1 ← {{0, 1,2, ...}}
```

```
for i1 = 1 to dp do
    for node in nodeSet1 do
    usedRows1 ← rowSet1[node1]
    for k1 = 1 to m1 do
        H1 ← new Histogram1()
        for j1 in usedRows1 do
            bin ← I1.f[k1][j1].bin
              H1[bin].y1 ← H1[bin].y1 + I.y1[j1]
              H1[bin].n1 ← H1[bin].n1+ 1
        Find the best split on histogram H1”.
```

Algorithm : GOSS Algorithm[29]

```
“Input: I1: training dataset, d1: iteration
Input: a1: sampling ratio of large gradient dataset
Input: b1: sampling ratio of small gradient dataset
Input: loss1: loss function, L1: weak learner
models ? {}, fact ? (1-a1)/b1
topN1 ? a1 × len(I1), randN1 ? b1 × len(I1)
for i1 = 1 to d1 do
    preds ? models.predict(I1) g1 ? loss(I1, preds), w1 ?
    {1, 1, ...}
    sorted ? GetSortedIndices(abs1(g1))
    topSet1 ? sorted[1:topN1]
    randSet1 ? RandomPick(sorted[topN1:len(I1)],
    randN1)
    usedSet1 ? topSet1 + randSet1
    w1[randSet1]× =fact .Assign weight fact to the
    small gradient data.
    newModel ? L1(I1[usedSet1], g1[usedSet1],
    W1[usedSet1])
    models.append(newModel1)”
```

3.3 Implementation of LightGBM

Application of Light GBM can be done easily. However the only intricacy is parameter modification. The technique of adjusting the essentials which regulator the behaviour of a given model is called parameter tuning. Light GBM covers more than 100 parameters. The basic classification of parameters are Control Parameters, Core Parameters, Metric Parameter, IO Parameter [29].

The proficiency of the model can be improved using the following methods:

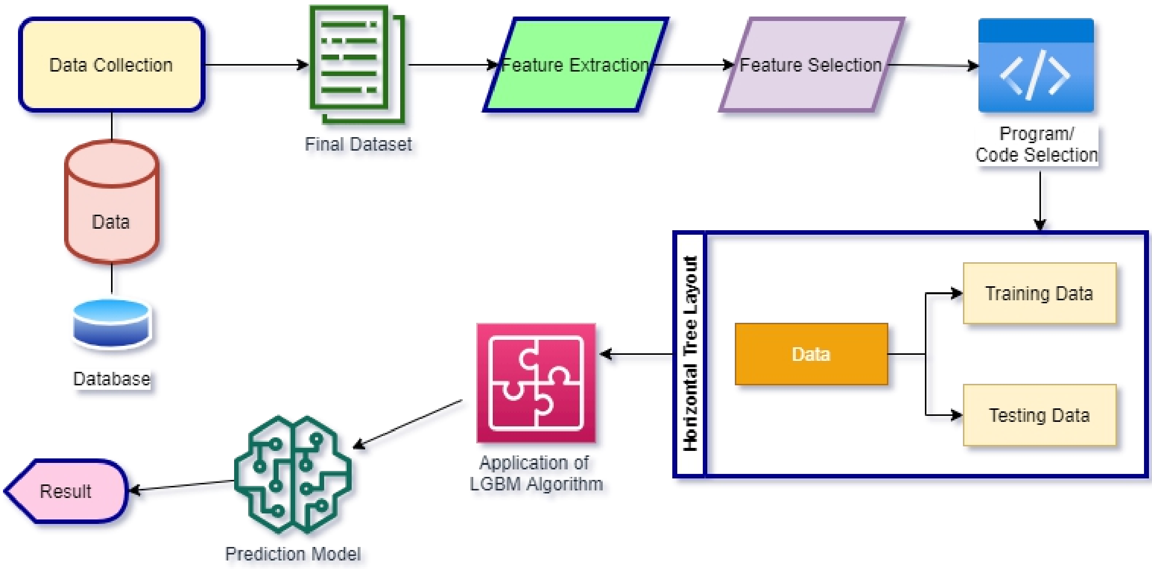
Number of leaves: The technique involved here is the numerates that is involved should be less than 2^(maximum depth) or equal to 2^(maximum depth). This number of leaves is considered as one of the main parameters as more cost will cause overfitting.

Minimum data stored in leaf: The data that is stored in the leaf should be of minimum cost. However, the cost that is set reflects on the size of the dataset that is involved.

Maximum depth: The depth of the tree can be estimated based on its maximum depth.

LightGBM is very fast in its processes with the existing gradient boosting trees implementations and due to its tree-growth technique, histogram-based memory and computation optimization. It consumes a complete solution for spread training of the dataset that is tested.[29]

4. Architecture



The construction of the analytical system starts from data collection unit as the base cluster and for predicting auto immune diseases, datasets containing historical medicinal data of a patient is significant. Such datasets can be obtained from databases such as Kaggle. Once such dataset that would be useful is PIMA Indian Dataset for diabetes and any other relevant datasets can be collected and consolidated to form the final dataset. The final dataset can nowadays be loaded into the IBM Watson cloud workspace running a related machine erudition instance. The machine learning instance studies the dataset and extracts the unique fields in dataset. This is called feature extraction.

Further, the extracted structures would now be available for user configuration. The user can now select the fields that need to be analyzed using a prognostic algorithm and the field which is the possible outcome of the analyzed fields. Any augmentations can be organized as well. This stage is the feature selection phase. The LGBM algorithm is used to train and test the dataset as it delivers maximum accuracy for the dataset tested. Once the algorithm is selected, the dataset can be divided into two splits for training and testing respectively. The calculation of data in dataset that needs to be used for training and testing can be customary by the user.

Once the training and testing split is set by the user, the algorithm will accordingly use the data in the dataset to train itself based on the outcome field present in the dataset and it will use the remaining split to test the accuracy of prediction. Once the "training and testing" is done, the model is ready to be used. This predictive model running the LGBM algorithm will now be ready to predict the possibility of a persistent having an autoimmune disease when a similar dataset containing applicable input fields is laden.

The user should keep in attention to gather all the patient details in streak with the dataset used to train and test the model, or else the prediction outcome would become inaccurate, as the model by itself is not 100% accurate in prediction.

Once the analysis is carried out by using the analytical model, the estimate results can be observed and extracted in a variety of set-ups such as xml, csv, spreadsheet, or word document as per the user's convenience. The results can be fine tuned by running the algorithm with any accessible developments and the accuracy of such enrichments can be pre-determined in the training and testing stage itself.

4. Results and Discussion

In this research studies, the PIMA Indian Dataset is used for testing. Algorithms such as Logistic Regression, XGB Classifier, Gradient Boosting Classifier, Decision Tree, Extra Trees Classifier, Random Forest, LGBM are used and a comparative study is thru to verify and analyze the accuracy of the algorithm. From this comparison it is resolved that LGBM Classifiers produce the best accuracy out of all the algorithms used and produces the desired results. The below table gives the accuracy of all the algorithms used.

Dataset	**Logistic Regression**	**XGB Classifier**	**Gradient Boosting Classifier**	**Decision Tree**	**Extra Trees Classifier**	**Random Forest**	**LGBM**
PIMA Indian Dataset	75.20%	83.30%	94.10%	94.40%	94.60%	94.80%	95.20%

5. Conclusion and Future Work

From the above done discussions it can be concluded that the LGBM Classifier algorithm can produce the most accurate result for the given PIMA Indian dataset. It is also beneficial to predict, detect and monitor the diabetic disease as per the prerequisite of the data model. In Future, this LGBM Algorithm can be executed with some progressive landscapes and an Advanced LGBM Algorithm can be castoff to bring out the necessary study.

Acknowledgement

This publication is a part of my PhD research program of **SRM Institute of Science and Technology, Chennai.**

Conflict of interest

The authors declared no conflict of interest.

Funding Support

The authors declared no funding support for this study.

References

1. Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. Journal of Big Data, Springer, 6, 54 (2019).

2. A. Ukil, S. Bandyoapdhyay, C. Puri and A. Pal, "IoT Healthcare Analytics: The Importance of Anomaly Detection," 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, 2016, pp. 994-997, doi: 10.1109/AINA.2016.158.

3. C. Sun, Q. Li, L. Cui, H. Li and Y. Shi, "Heterogeneous network-based chronic disease progression mining," in Big Data Mining and Analytics, vol. 2, no. 1, pp. 25-34, March 2019, doi: 10.26599/BDMA.2018.9020009.

4. Himansu Das, Bighnaraj Naik, H.S. Behera, Corrigendum “Medical disease analysis using neuro-fuzzy with feature extraction model for classification” [Inf Med Unlocked 18 (2020) 1–12 page/100288], Informatics in Medicine Unlocked, Volume 18, 2020, Pages 100299.

5. O. Kolesnichenko et al., "Big Data Analytics of Inpatients Flow with Diabetes Mellitus type 1 : Revealing new awareness with Advanced Visualization of Medical Information System Data," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 191-196, doi: 10.1109/CONFLUENCE.2019.8776910.

6. Bikku, T. Multi-layered deep learning perceptron approach for health risk prediction. J Big Data 7, Springer, 50 (2020),https://doi.org/10.1186/s40537-020-00316-7.

7. Nibareke, T., Laassiri, J. Using Big Data-machine learning models for diabetes prediction and flight delays analytics. J Big Data 7, Springer, 78 (2020). https://doi.org/10.1186/s40537-020-00355-0.

8. Hosseini, M.M., Zargoush, M., Alemi, F. et al. Leveraging machine learning and big data for optimizing medication prescriptions in complex diseases: a case study in diabetes management. J Big Data 7, Springer, 26 (2020). https://doi.org/10.1186/s40537-020-00302-z.

9. Huang, Z., Dong, W., Bath, P. et al. On mining latent treatment patterns from electronic medical records. Data Min Knowl Disc 29, Springer, 914–949 (2015). https://doi.org/10.1007/s10618-014-0381-y.

10. Wang, F., Stiglic, G., Obradovic, Z. et al. Guest editorial: Special issue on data mining for medicine and healthcare. Data Min Knowl Disc 29, Springer, 867–870 (2015). https://doi.org/10.1007/s10618-015-0414-1.

11. Goyal, J., Khandnor, P. & Aseri, T.C. A Comparative Analysis of Machine Learning classifiers for Dysphonia-based classification of Parkinson’s Disease. Int J Data Sci Anal 11, Springer,69–83 (2021). https://doi.org/10.1007/s41060-020-00234-0.

12. S. Kumar and M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools," in Big Data Mining and Analytics, vol. 2, no. 1, pp. 48-57, March 2019, doi: 10.26599/BDMA.2018.9020031.

13. Strang, K.D. Problems with research methods in medical device big data analytics. Int J Data Sci Anal 9, Springer, 229–240 (2020). https://doi.org/10.1007/s41060-019-00176-2.

14. Mercaldo, F.; Nardone, V.; Santone, A. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Comput. Sci. 2017, 112, 2519–2528.

15. Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 2018, 132, 1578–1585.

16. Kandhasamy, J.P.; Balamurali, S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput. Sci. 2015, 47, 45–51.

17. Tafa, Z.; Pervetica, N.; Karahoda, B. An intelligent system for diabetes prediction. In Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 14–18 June 2015; pp. 378–382.

18. Yuvaraj, N.; SriPreethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust. Comput. 2017, 22, 1–9.

19. Olaniyi, E.O.; Adnan, K. Onset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng. Res. 2014,5, 754–759.

20. Soltani, Z.; Jafarian, A. A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II. Int. J. Adv. Comput. Sci. Appl. 2016, 7, 89–94.

21. Somnath, R.; Suvojit, M.; Sanket, B.; Riyanka, K.; Priti, G.; Sayantan, M.; Subhas, B. Prediction of Diabetes Type-II Using a Two-Class Neural Network. In Proceedings of the 2017 International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; pp. 65–7.

22. Mamuda, M.; Sathasivam, S. Predicting the survival of diabetes using neural network. In Proceedings of the AIP Conference Proceedings, Bydgoszcz, Poland, 9–11 May 2017; Volume 1870, pp. 40–46.

23. Negi, A.; Jaiswal, V. A first attempt to develop a diabetes prediction method based on different global datasets. In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC),Waknaghat, India, 22–24 December 2016; pp. 237–241.

24. John Martinsson, Alexander Schliep, Björn Eliasson, Olof Mogren, Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks, March 2020, Springer, Journal of Healthcare Informatics Research 4(2), DOI: 10.1007/s41666-019-00059-y

25. Jayanthi, N., Vijaya Babu, B., & Sambasiva Rao, N., Survey on clinical prediction models for diabetes prediction, Journal of Big Data volume 4, Article number: 26 (2017).

26. Guolin Ke , Qi Meng , Thomas Finley , Taifeng Wang , Wei Chen , Weidong Ma1 , Qiwei Ye , Tie-Yan Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

27. Ambigavathi, M., Sridharan, D..Big Data Analytics in Healthcare, IEEE Tenth International Conference on Advanced Computing (ICoAC), December 2018; pp-269-276.

28. Souad Larabi-Marie-Sainte , Linah Aburahmah , Rana Almohaini and Tanzila Saba, Current Techniques for Diabetes Prediction: Review and Case Study, Applied Sciences, MDPI Journal,14 September 2019; Accepted: 21 October 2019; Published: 29 October 2019.

29. O. Kolesnichenko et al., "Big Data Analytics of Inpatients Flow with Diabetes Mellitus type 1 : Revealing new awareness with Advanced Visualization of Medical Information System Data," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 191-196, doi: 10.1109/CONFLUENCE.2019.8776910.

30. N. Magdelaine et al., "A Long-Term Model of the Glucose–Insulin Dynamics of Type 1 Diabetes," in IEEE Transactions on Biomedical Engineering, vol. 62, no. 6, pp. 1546-1552, June 2015, doi: 10.1109/TBME.2015.2394239.

31. Zhu, T., Li, K., Chen, J. et al. Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes. J Healthc Inform Res 4, Springer, 308–324 (2020). https://doi.org/10.1007/s41666-020-00068-2