2024 3rd International Conference for Innovation in Technology (INOCON)
Karnataka, India. Mar 1-3, 2024

Using Machine Learning-based SMOTE Analysis with the Light GBM Classification Method to Classify Diabetic Patients

Kanwarpartap Singh Gill [1]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
kanwarpartap.gill@chitkara.edu.in

Vatsala Anand [2]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
vatsala.anand@chitkara.edu.in

Rahul Chauhan[3]
Computer Science & Engineering, Graphic Era Hill University,
Dehradun, Uttarakhand, India, 248002
chauhan14853@gmail.com

Hemant Singh Pokhariya[4]
Computer Science & Engineering, Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India, 248002
hemantdoon86@gmail.com

2024 3rd International Conference for Innovation in Technology (INOCON) | 979-8-3503-8193-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INOCON60754.2024.10511914

***Abstract*— Diabetes is a prevalent and enduring metabolic condition that impacts a significant global population. The timely identification and categorization of this ailment are of paramount importance in order to provide optimal control and therapeutic interventions. Machine learning algorithms have shown encouraging outcomes in the classification of diabetes using patient data. The gradient boosting framework known as LightGBM is often favoured for this particular job because to its notable efficiency and high level of predicted accuracy. Moreover, the Synthetic Minority Over-sampling Technique (SMOTE) may be used to mitigate the issue of class imbalance in the diabetes dataset, a prevalent challenge encountered in the study of medical data. The primary objective of this work is to use the LightGBM classifier for the purpose of classifying the diabetes condition. The objective will be accomplished by the use of SMOTE Analysis, which aims to tackle the challenge of imbalanced data. Subsequently, an examination of both balanced and unbalanced datasets will be conducted, leading to the attainment of a 72 percent accuracy rate. Future researchers in this specific field of inquiry will benefit from the significant level of accuracy shown in this study.**

Keywords*— *Artificial Intelligence, Deep Learning, Diabetes Classification Analysis, Model Training, Classification, LightGBM Classifier, SMOTE, Deep Learning

I. INTRODUCTION

This research study investigates the use of LightGBM in the classification of diabetes, combining the analysis of Synthetic Minority Over-sampling Technique (SMOTE) to enhance the performance of the model.

The categorization of diabetes using LightGBM is achieved by using tree-based learning techniques inside the framework. LightGBM is a gradient boosting framework specifically designed for classification and regression applications. The prevalence of this phenomenon has been attributed to its notable attributes of efficiency, scalability, and precision. LightGBM, an advanced machine learning algorithm, may be effectively used in the context of diabetes classification. By leveraging a range of patient characteristics, including age, gender, BMI, blood pressure, and glucose levels, LightGBM can accurately forecast the presence or absence of diabetes in a given patient.

The problem of class imbalance is often seen in medical datasets, such as those including diabetes data, when the prevalence of non-diabetic patients tends to surpass that of diabetic cases. The Synthetic Minority Over-sampling Technique (SMOTE) is a method used to address class imbalance in datasets. By generating synthetic instances of the minority class, SMOTE aims to equalise the distribution of classes, hence reducing the bias towards the majority class in predictive models. By using the Synthetic Minority Over-sampling Technique (SMOTE) on the diabetes dataset, the model's capacity to accurately identify instances of diabetes may be enhanced.

Feature selection plays a crucial role in the development of a diabetic classification model. It is important to identify pertinent variables, such as glucose levels, body mass index (BMI), and family medical history, while excluding irrelevant variables. Moreover, the process of feature engineering encompasses the creation of novel features as well as the transformation of pre-existing ones, with the aim of augmenting the predictive capabilities of the model.

The effectiveness of LightGBM enables expedited model training in comparison to alternative gradient boosting frameworks. The optimisation of a model's performance is crucially dependent on the process of hyperparameter tweaking. In order to optimise performance, it is necessary to carefully adjust parameters such as learning rate, number of estimators, and maximum depth of trees.

Cross-validation and evaluation approaches are used to ascertain the model's capacity to generalise. One such technique is k-fold cross-validation. The performance of the model may be assessed by using many measures, including accuracy, precision, recall, F1-score, and ROC-AUC.

The act of interpreting and providing explanations for the predictions made by a model is of utmost importance within the medical field. Techniques such as SHAP (SHapley Additive exPlanations) might facilitate the comprehension of the decision-making process used by the model, hence enhancing its transparency and reliability.

After the development of a strong diabetes classification model based on LightGBM, it may be used in clinical environments to aid healthcare professionals in the diagnosis of diabetes, facilitating treatment choices, and enhancing patient outcomes.

II. LITERATURE

A low code development strategy was used by Whig, P. et al. The Pycaret machine learning methodology is used for the goal of classifying, identifying, and forecasting diabetes [1-2]. The aim of the research undertaken by Monteiro - Soares, M. et.al. was to assess previously published and

utilised techniques for the characterisation of ulcers in patients with diabetes. The objective of the study was to determine the optimal system that could be suggested for diverse applications, such as enhancing efficient communication among healthcare practitioners, forecasting the clinical prognosis of individual ulcers, profiling individuals with infection and/or peripheral arterial disease, and conducting comparative analyses of outcomes across distinct populations [3-4]. Omar et al. undertook a thorough investigation of previous scholarly studies pertaining to the classification of diabetes into discrete subtypes. The researchers not only conducted a comprehensive examination and analysis of these investigations, but also presented a succinct overview of their outcomes. Furthermore, the scholars used data mining methodologies and several clustering algorithms to proficiently implement the procedure of diabetes subtyping [5-6]. Chang et al. presented a research proposal for an electronic diagnostic system that incorporates machine learning (ML) methodologies. The purpose of this system is to be implemented within the framework of the Internet of Medical Things (IoMT), with a particular emphasis on the identification of type 2 diabetes, medically referred to as diabetes mellitus [7-8]. Research was undertaken by ElSayed, N.A. and colleagues to investigate the significance of offering counselling services to persons who have positive results for autoantibodies. The counselling session should prioritise the dissemination of information regarding the potentiality of developing diabetes, the various symptoms associated with diabetes, strategies for preventing diabetic ketoacidosis (DKA), and the potential need for further diagnostic testing to ascertain eligibility for interventions aimed at impeding the advancement of the disease [9-10]. Monteiro - Soares et al. conducted a thorough compilation of categorization systems that had the potential to be effectively used in a clinical setting. The compilation presented in this study was derived from an evaluation of diagnostic tests, focusing particularly on the assessment of usability, accuracy, and reliability of each system in predicting problems associated with ulcers, along with their resource utilisation [11-12]. Leslie, R.D. et al. presented a thorough exposition on the diversity of diabetes, highlighting current approaches that may improve the treatment of the condition. The methodologies used in this study include the incorporation of three distinct disease models, namely the palette model, the threshold model, and the gradient model, which together comprise several forms of diabetes [13-14]. The research conducted by Almutairi, E.S. and colleagues aimed to assess the effectiveness of various categorization methods in accurately classifying the prevalence rates of diabetes, as well as predicting the patterns of the disease based on associated behavioural risk factors like smoking, obesity, and physical inactivity. This investigation specifically focused on the context of Saudi Arabia, as indicated by the relevant references [15-16]. Kalyani et al. introduced an improved capsule network design with the objective of identifying and categorising instances of diabetic retinopathy [17-18]. Thotad, P.N. and colleagues used machine learning methodologies to discern and classify diabetes cases of notable clinical relevance [19-20].

The main aim of this study is to use the LightGBM Classifier in order to classify diabetes throughout the evaluation phase.

The assessed parameters in this study exhibit potential in assisting researchers who have faced difficulties in developing a strong organisational framework.

The debate encompasses several topics, including strategies for enhancement and their subsequent impacts on outcomes.

The study encompasses several subtopics. The subsequent piece of the research paper offers a comprehensive elucidation of the dataset used for analysis, while the adjoining segment delineates the Anova Test for feature selection and the methodologies employed to tackle the structure of the LightGBM Classifier. Section 6 of the presentation contains the provided results, while Section 7, found near the conclusion of the presentation, includes a detailed compilation of references.

III. INPUT DATASET

The dataset under consideration is derived from the National Institute of Diabetes and Digestive and Kidney Diseases. The main objective of the dataset is to effectively predict the occurrence or non-occurrence of diabetes in individuals, using certain diagnostic characteristics that are included within the dataset. Several constraints were applied to the selection process of these instances from a larger database. The research focuses only on adult females who are at least 21 years old and belong to the Pima Indian ethnic group. The dataset used in this investigation was obtained from Kaggle and consists of several medical predictor elements together with a solitary objective variable, referred to be Outcome (as shown in Fig. 1). The predictor variables include several aspects, including the patient's number of pregnancies, body mass index (BMI), insulin level, age, and other pertinent factors.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.60	0.63	50	1
1	1	85	66	29	0	26.60	0.35	31	0
2	8	183	64	0	0	23.30	0.67	32	1
3	1	89	66	23	94	28.10	0.17	21	0
4	0	137	40	35	168	43.10	2.29	33	1

Fig. 1. Extraction of Features from a Dataset in CSV Format

IV. THE ANALYSIS OF VARIANCE (ANOVA) TEST FOR FEATURE SELECTION IN NUMERICAL VARIABLES

Feature selection has great significance in the domains of machine learning and statistical analysis. The fundamental aim of this task is to identify the most relevant features within a certain dataset, while concurrently excluding those that provide less informative significance. The Analysis of Variance (ANOVA) test is a statistically rigorous approach used to ascertain the numerical elements that have a significant impact on the target variable. The use of the analysis of variance (ANOVA) test is a very valuable approach for performing feature selection in datasets including numerical attributes and a categorical target variable.

The analysis of variance (ANOVA) facilitates the assessment of the relative significance of each feature's impact on the target variable, as seen in Fig. 2. This statistical methodology enables the improvement of the feature set, hence boosting the effectiveness and interpretability of machine learning models. A comprehensive analysis of significance levels and the

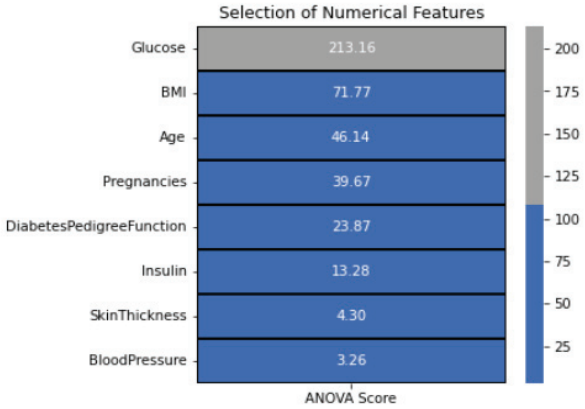

Fig. 2. ANOVA test for selection of Numerical Features

assessment of ANOVA assumptions are essential for the successful use of feature selection. Based on the findings of the analysis of variance (ANOVA) test, it can be deduced that there is a positive relationship between the ANOVA score and the importance of the feature under investigation.

V. LIGHT GBM CLASSIFIER FOR CLASSIFICATION OF DIABETES

LightGBM is a widely acclaimed gradient boosting framework that has garnered significant attention in the area of machine learning because to its robustness and efficiency. It has been widely used for a multitude of applications, including classification. LightGBM has significant benefits in the categorization of diabetes, owing to its exceptional speed, high accuracy, and capacity to effectively handle extensive datasets. LightGBM is a machine learning algorithm that utilises the gradient boosting approach, a kind of ensemble learning. Ensemble learning is a technique that integrates numerous weak learners, often decision trees, in order to construct a robust predictive model. Within the domain of diabetes classification, LightGBM employs ensemble learning techniques to enhance the precision and resilience of the classifier as shown in Fig. 3.

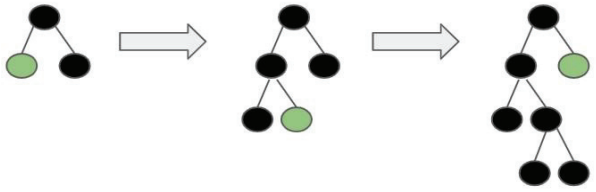

Fig. 3. LightGBM Classifier architecture

In summary, the rapidity, precision, and capacity of LightGBM in managing extensive datasets provide it a tempting option for the categorization of diabetes. When used with appropriate feature selection, hyperparameter tuning, and interpretability methodologies, the utilisation of LightGBM has the potential to provide precise and clinically significant predictions, therefore augmenting the process of diagnosing and treating diabetes.

VI. RESULTS

A. Classification Report for Diabetes Prediction Using the LightGBM Classifier on an Unbalanced Dataset

The prediction of diabetes is a crucial undertaking within the healthcare domain; nevertheless, it often necessitates handling imbalanced datasets characterised by a substantial disparity between the number of non-diabetic patients and diabetic cases. Machine learning techniques, such as the LightGBM classifier, provide a robust approach to address the issue of imbalanced data. This brief theoretical analysis delves into the importance of a classification report in the context of using the LightGBM classifier for the prediction of diabetes on datasets that exhibit an imbalance in class distribution as shown in Fig. 4.

	precision	recall	f1-score	support
0	0.78	0.84	0.81	134
1	0.54	0.45	0.49	58
accuracy			0.72	192
macro avg	0.66	0.64	0.65	192
weighted avg	0.71	0.72	0.71	192

Fig. 4. LightGBM Classifier Classification Report on an Unbalanced Dataset

In summary, the use of a classification report seems to be a helpful resource in the assessment of the LightGBM classifier's efficacy in predicting diabetes, particularly in the context of imbalanced datasets. The use of this tool enables healthcare practitioners and data scientists to make well-informed judgements about the appropriateness of the model for early diagnosis and patient risk assessment, while taking into account the difficulties presented by class imbalance.

B. Illustration of the Diabetes Prediction Confusion Matrix using an Unbalanced Dataset and the LightGBM Classifier

In the domain of diabetes classification, machine learning algorithms often face imbalanced datasets characterised by a significant disparity in the frequency of non-diabetic cases compared to diabetic patients. The use of a confusion matrix is of great significance in evaluating the efficacy of the LightGBM classifier on datasets of this kind. This brief theoretical analysis examines the significance of the confusion matrix as a visual aid for comprehending the classification outcomes in the context of diabetes prediction while using LightGBM as shown in Fig. 5.

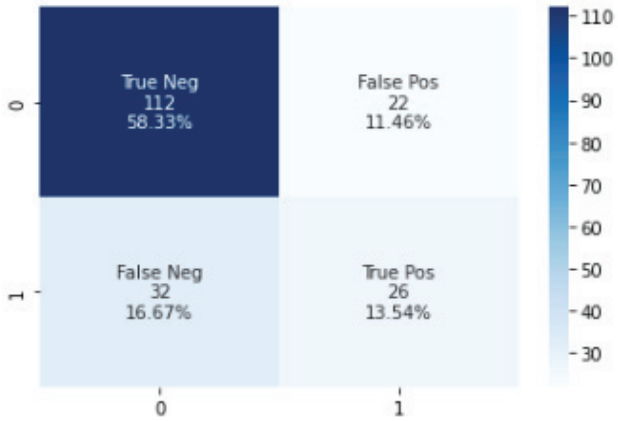

Fig. 5. Illustration of the Confusion Matrix of the LightGBM Classifier for Unbalanced Data

C. *Classification Report of a LightGBM Classifier on a Balanced Dataset Attempting to Predict Diabetes*

The importance of balanced datasets in machine learning cannot be overstated, especially when it comes to tasks like as diabetes classification. Class imbalances within the dataset may significantly impact the performance of the model, introducing biases that may compromise its accuracy and reliability. The use of a classification report is a helpful tool in the assessment of the effectiveness of the LightGBM classifier in the categorization of diabetes, especially in scenarios when datasets are balanced. This tool enables healthcare practitioners and data scientists to make well-informed judgements about the appropriateness of the model for early diagnosis and patient risk assessment. It ensures that the model functions well, while mitigating any biases that may arise due to class imbalance as shown in Fig. 6.

	precision	recall	f1-score	support
0	0.75	0.75	0.75	114
1	0.79	0.79	0.79	136
accuracy			0.78	250
macro avg	0.77	0.77	0.77	250
weighted avg	0.78	0.78	0.78	250

Fig. 6. Report on the LightGBM Classifier's Classification of a Neutral Dataset

D. *Balanced Dataset Confusion Matrix Illustration for the Classification of Diabetes Employing the LightGBM Classifier*

The utilisation of balanced datasets is critical for the precise evaluation and forecasting of models within the domain of diabetes classification. Employing a confusion matrix to visualise and comprehend the performance of a machine learning model, such as the LightGBM classifier, on a balanced dataset in diabetes classification is highly effective as shown in Fig. 7.

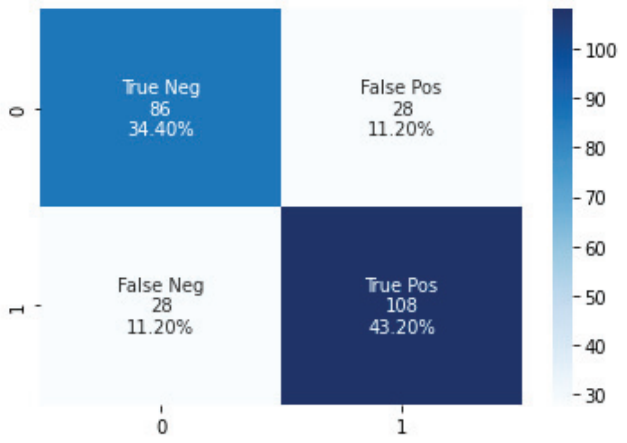

Fig. 7. Balanced Data Confusion Matrix Illustration for LightGBM Classifier

VII. Conclusion

In summary, the utilisation of the LightGBM classifier for diabetes prediction signifies a robust and efficacious methodology in the realm of early detection and evaluation of threats. In the healthcare industry, LightGBM has become a valuable instrument due to its effectiveness, precision, and capability to process balanced and unbalanced datasets. By means of patient data analysis and the application of sophisticated machine learning methodologies, LightGBM is capable of providing healthcare professionals with precise predictions and invaluable insights. The evaluation process yields classification reports and confusion matrices that furnish a comprehensive assessment of the model's performance. These metrics guarantee that the model strikes a delicate equilibrium between accurately identifying true positive diabetic cases and reducing the occurrence of false positives and false negatives. Ensuring this equilibrium is vital within the medical field, where the reduction of superfluous interventions and early detection are both critical. Progress in the domain of diabetes prognosis necessitates ongoing refinement and optimisation of models such as LightGBM, with the integration of interpretability and explainability serving to foster confidence among patients and healthcare practitioners. This facilitates early diagnosis, provides guidance for treatment decisions, and ultimately contributes to enhanced patient outcomes. LightGBM's application in diabetes prediction exemplifies the capacity of machine learning to revolutionise the healthcare industry, rendering it more proactive, streamlined, and patient-focused amidst an era of escalating healthcare data. By means of ongoing investigation, enhancement, and incorporation of cutting-edge methodologies, it is possible to anticipate diabetes prediction models that are even more precise and dependable, thereby augmenting the quality of life for individuals afflicted with this chronic ailment.

References

[1] Whig, P., Gupta, K., Jiwani, N., Jupalle, H., Kouser, S. and Alam, N., 2023. A novel method for diabetes classification and prediction with Pycaret. Microsystem Technologies, pp.1-9.

[2] Bharany, S., Badotra, S., Sharma, S., Rani, S., Alazab, M., Jhaveri, R.H. and Gadekallu, T.R., 2022. Energy efficient fault tolerance techniques in green cloud computing: A systematic survey and taxonomy. Sustainable Energy Technologies and Assessments, 53, p.102613.

[3] Monteiro‐Soares, M., Hamilton, E.J., Russell, D.A., Srisawasdi, G., Boyko, E.J., Mills, J.L., Jeffcoate, W. and Game, F., 2023. Classification of foot ulcers in people with diabetes: a systematic review. Diabetes/Metabolism Research and Reviews, p.e3645.

[4] Sharma, B. and Koundal, D., 2018. Cattle health monitoring system using wireless sensor network: a survey from innovation perspective. IET Wireless Sensor Systems, 8(4), pp.143-151.

[5] Omar, N., Nazirun, N.N., Vijayam, B., Wahab, A.A. and Bahuri, H.A., 2023. Diabetes subtypes classification for personalized health care: A review. Artificial Intelligence Review, 56(3), pp.2697-2721.

[6] Singh, P.K. and Sharma, A., 2022. An intelligent WSN-UAV-based IoT framework for precision agriculture application. Computers and Electrical Engineering, 100, p.107912.

[7] Chang, V., Bailey, J., Xu, Q.A. and Sun, Z., 2023. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), pp.16157-16173.

[8] Reshan, M.S.A., Gill, K.S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A. and Shaikh, A., 2023, May. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.

[9] ElSayed, N.A., Aleppo, G., Aroda, V.R., Bannuru, R.R., Brown, F.M., Bruemmer, D., Collins, B.S., Hilliard, M.E., Isaacs, D., Johnson, E.L. and Kahan, S., 2023. Addendum. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. Diabetes Care 2023; 46 (Suppl. 1): S19–S40. Diabetes care, 46(9), pp.1715-1715.

[10] Gill, K.S., Anand, V. and Gupta, R., 2023, August. An Efficient VGG19 Framework for Malaria Detection in Blood Cell Images. In

2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-4). IEEE.

[11] Monteiro‐Soares, M., Hamilton, E.J., Russell, D.A., Srisawasdi, G., Boyko, E.J., Mills, J.L., Jeffcoate, W. and Game, F., 2023. Guidelines on the classification of foot ulcers in people with diabetes (IWGDF 2023 update). Diabetes/Metabolism Research and Reviews, p.e3648.

[12] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, May. Smart Shoe Classification Using Artificial Intelligence on EfficientnetB3 Model. In 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 254-258). IEEE.

[13] Leslie, R.D., Ma, R.C.W., Franks, P.W., Nadeau, K.J., Pearson, E.R. and Redondo, M.J., 2023. Understanding diabetes heterogeneity: key steps towards precision medicine in diabetes. The Lancet Diabetes & Endocrinology.

[14] Yang, M., Kumar, P., Bhola, J. and Shabaz, M., 2021. Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. International Journal of System Assurance Engineering and Management, pp.1-9.

[15] Almutairi, E.S. and Abbod, M.F., 2023. Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia. Modelling, 4(1), pp.37-55.

[16] Sharma, S., Gupta, S., Gupta, D., Juneja, S., Gupta, P., Dhiman, G. and Kautish, S., 2022. Deep learning model for the automatic classification of white blood cells. Computational Intelligence and Neuroscience, 2022.

[17] Kalyani, G., Janakiramaiah, B., Karuna, A. and Prasad, L.N., 2023. Diabetic retinopathy detection and classification using capsule networks. Complex & Intelligent Systems, 9(3), pp.2651-2664.

[18] Sharma, R. and Kukreja, V., 2022, March. Amalgamated convolutional long term network (CLTN) model for Lemon Citrus Canker Disease Multi-classification. In 2022 International conference on decision aid sciences and applications (DASA) (pp. 326-329). IEEE.

[19] Thotad, P.N., Bharamagoudar, G.R. and Anami, B.S., 2023. Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 17(1), p.102690.

[20] Sasubilli, S.M., Kumar, A. and Dutt, V., 2020, June. Machine learning implementation on medical domain to identify disease insights using TMS. In 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 1-4). IEEE.