DATA MINING AND ENSEMBLE TECHNIQUES FOR IMPROVED DIABETIC CLASSIFICATION WITH MACHINE LEARNING

K. SREEDEVI
Research Scholar
Dept of Computer science and Engineering
NIILM University- Haryana.

Dr. Anshul Mishra
Research Supervisor
Dept of Computer science and Engineering
NIILM University- Haryana.

ABSTRACT

Diabetes, also known as chronic illness, is a group of metabolic diseases due to a high level of sugar in the blood over a long period. The risk factor and severity of diabetes can be reduced significantly if the precise early prediction is possible. Diabetic retinopathy is identified by red spots known as microanuerysms and bright yellow lesions called exudates. It has been observed that early detection of exudates and microaneurysms may save the patient's vision and this paper proposes a simple and effective technique for diabetic retinopathy. An automated approach that uses image processing, features extraction and machine learning models to predict accurately the presence of the exudates and micro aneurysms which can be used for grading has been proposed. The robust and accurate prediction of diabetes is highly challenging due to the limited number of labeled data and also the presence of outliers in the diabetes datasets. In this literature, we are proposing a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Naive Bayes, and XGBoost) to improve the prediction of diabetes where the weights are estimated from the corresponding Area Under ROC Curve (AUC) of the ML model. AUC is chosen as the performance metric, which is then maximized during hyperparameter tuning using the grid search technique.

Keywords: *Diabetic retinopathy, Machine Learning (ML), k-nearest Neighbour, image processing, ML model.*

INTRODUCTION

Diabetes datasets play a crucial role in early-stage detection. Mentioned in the article that patient record is obtained from electronic devices and paper records, with the automotive record having timestamps event and the paper record having only "logical time" slots. The diabetes datasets help to detect diabetes in a person. Attributes of a dataset play a crucial role in detection. Researchers and scientists use different data files to predict diabetes. The ML-based model is required for relevant datasets having essential aspects for training and validations. Choosing the appropriate elements from the dataset increased the ability of the ML model to predict accurate data. The timely identification of type 2 diabetes holds significance due to its potential to significantly mitigate long-term complications through rigorous diabetes management. Nevertheless, conducting diabetes screening across the entire population lacks cost-effectiveness, thus emphasizing the need to prioritize the recognition of individuals with a heightened susceptibility to the condition. Numerous investigations regarding diabetes screening have been conducted within the previous ten years. Risk prediction or stratification models can serve the purpose of identifying individuals at an elevated risk level for diabetes, allowing for subsequent targeted testing. Typically, these models incorporate a blend of variables, encompassing weight, lifestyle, familial background, and clinical measurements, and are formulated through the utilization

of multivariable statistical techniques. Nevertheless, numerous of these models are not extensively employed within clinical practice, primarily owing to their foundation on data gathered for alternate objectives. This circumstance can decrease the relevance of these findings when applied to a broader population. Additionally, attempts are often made to create models that are easy to use in clinical practice. This is often accomplished by condensing continuous variables into distinct categories or opting for predictors in a subjective manner. However, such approaches can result in excessive simplification and a consequent decrease in the models' overall efficacy.

LITERATURE REVIEW

Saad Althobaiti (2024) Diabetes Mellitus (DM) is an enduring metabolic illness that disturbs many individuals globally. This study addresses the global impact of Diabetes Mellitus (DM) and emphasizes the critical role of accurate DM detection in early diagnosis, effective treatment, and prevention of complications. The research introduces an optimized DM detection model, the GBM-DRU (Gradient Boosting Machine - Data Reduction Unit) network, which integrates feature engineering and ensemble learning techniques to enhance prediction accuracy and support clinical decision-making. The GBM-DRU network combines the powerful gradient boosting machine algorithm with a data reduction unit (DRU) for efficient feature selection, reducing dimensionality and improving computational efficiency. Feature engineering enhances discriminatory power, while ensemble learning methods, including bagging and boosting, improve overall model performance.

Mamta Bansal (2023) Diabetes is the leading cause of death in the world, and it also affects kidney disease, loss of vision, and heart disease. Data mining techniques contribute to health care decisions for accurate disease diagnosis and treatment, reducing the workload of experts. Diabetes prediction is a rapidly expanding field of research. Early diabetes prediction will result in improved treatment. Diabetes causes a variety of health issues. Therefore, it is critical to prevent, monitor, and raise awareness about it. In this study, we propose a diabetes prediction model using data mining techniques. We apply four data mining techniques such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. The proposed mechanism is trained using Python and analysed with a real dataset, which is collected from Kaggle. Furthermore, the performance of the proposed mechanism is analysed using the confusion matrix, sensitivity and accuracy performance metrices. In logistic regression, the accuracy is high, i.e., 82.46%, in comparison to other data mining techniques.

Pijush Kanti Dutta Pramanik (2023) Diabetes is considered one of the leading healthcare concerns affecting millions worldwide. Taking appropriate action at the earliest stages of the disease depends on early diabetes prediction and identification. To support healthcare providers for better diagnosis and prognosis of diseases, machine learning has been explored in the healthcare industry in recent years. The results were analysed using various statistical/machine learning metrics and k-fold cross-validation techniques. Gradient boosting achieved the greatest accuracy rate of 92.85% among all the classifiers.

Precision, recall, f1-score, and receiver operating characteristic (ROC) curves were used to further validate the model. The suggested model outperformed the current studies in terms of prediction accuracy, demonstrating its applicability to other diseases with similar predicate indications.

Worku Gachena Negera (2021) Diabetes mellitus (DM) is a severe chronic disease that affects human health and has a high prevalence worldwide. Research has shown that half of the diabetic people throughout the world are unaware that they have DM and its complications are increasing, which presents new research challenges and opportunities. In this paper, we propose a preemptive diagnosis method for diabetes mellitus (DM) to assist or complement the early recognition of the disease in countries with low medical expert densities. Diabetes data are collected from the Zewditu Memorial Hospital (ZMHDD) in Addis Ababa, Ethiopia. Light Gradient Boosting Machine (LightGBM) is one of the most recent successful research findings for the gradient boosting framework that uses tree-based learning algorithms.

Beschi Raja (2019) Diabetes is one of the most common diseases for both adults and children. Machine Learning Techniques help to identify the disease in an earlier stage to prevent it. This work presents the effectiveness of the Gradient Boosted Classifier which is unfocused in earlier existing works. It is compared with two machine learning algorithms such as Neural Networks, Radom Forest employed on benchmark Standard UCI Pima Indian Dataset. The models created are evaluated by standard measures such as AUC, Recall, and Accuracy. As expected, the Gradient boosted classifier outperforms the other two classifiers in all performance aspects.

Machine Learning

Artificial Intelligence (AI) is becoming more mainstream in today's technologically advanced society. AI uses rules-based algorithms to boost machine intelligence. Machine learning is a fast-expanding area of AI with many practical applications in finance, healthcare, retail, data security, autonomous vehicles, image processing, computer vision, and more. Presently, these machine learning algorithms are used in virtually every aspect of the digital world. Data has multiplied throughout time, making it critical to track it to make important choices efficiently. For this purpose, machine learning techniques are indispensable. The usage of machine learning algorithms is common in data mining, which refers to analysing large amounts of data to uncover hidden links and patterns from big commercial databases containing lists of important records.

Finding this pattern helps one to anticipate or focus on critical information to solve an issue. Hence, machine learning has a hype in the rapidly expanding discipline of data science. An English computer scientist, and mathematician, devised the "Turing Test" to evaluate if a computer is intelligent. To succeed, a computer must fool a human into thinking it is human as well.

A prominent figure in the field of machine learning, coined the phrase "machine learning". As an employee at "IBM" in 1959, he created a software that became proficient in defeating him in the game of checkers. Tom Mitchell, another machine learning specialist, provided a thorough formulation in 1998 with a Well-posed Learning Problem: Learning by a computer program occurs when their ability to perform a task T, as defined by the symbol P, improves over time due to accumulated

experience E. Over the years, several breakthroughs have used artificial intelligence. As time has shown, these innovations outperform humans.

Supervised Learning

Supervised learning involves feeding machines labelled data, where the inputs are independent features, and the goal output is dependent. Humans tell the machine the input (typically vectors) and the expected result. In supervised learning, a "teacher" provides a training set of (X, Y) pairs. So, there is a labelled training data, using which a function is derived. A collection of training examples constitutes the training data. A supervised learner attempts to forecast the desired output of a function using input items. This prediction is based on the provided training examples. Two main types of problems can be solved using supervised learning: classification (prediction of object's class) and regression (having a continuous output). Again, for each of these problems, further division of algorithms exist, which are linear regression, logistic regression, k-nearest neighbour, support vector machine (SVM), decision tree and many more. The diagram illustrates the mechanism behind supervised learning

Brief of diabetes mellitus

Diabetes Mellitus occurs in humans and animals. Diabetes affects 1.5 million people in a year. This disease occurs when the disorder of carbohydrate metabolism shows the character of impaired ability in the human body not producing enough quantity of insulin. The supply of insulin level is not maintained in the human body, as mentioned in the study. Diabetes results from a malfunctioning pancreas in the human body, as mentioned in the study. For example, when the human pancreas is not producing insulin as it should. The human body does not supply insulin adequately. A person may get affected by diabetes due to three main reasons: genetics, surroundings, and an affluent lifestyle. Lifestyle is an essential factor in becoming a person diabetic and non-diabetic. If a person's ancestor is non-diabetic previously but the person has diabetes, then this has to be done by an affluent lifestyle. The next cause is the environmental impact. Now, people are influenced by the effects of losing weight. To lose weight people are taken poor diet and frequent exercises. That causes the problem of renal failure. Renal failure is the first step towards post-diabetes. In this modern world, people are facing issues like frequent hunger, frequent urination, weight loss and vision loss due to these signs of diabetes. The benevolence of this review paper in the domain of diabetic research considered ML-based access in DM detection, prevention, self-management, and personification. A review paper plays a crucial role in the study because it efficiently summarizes cutting-edge research in a specific area.

Pregestational diabetes

Pregestational diabetes is the combination of Type-1 and Type-2 diabetes that occurs in females during the time of pregnancy. Pregnant women have pre-existing diabetes during the time of pregnancy. According to pregestational increases the risk of malfunctions in the baby during the growth process in the ovary. The result of this type of diabetes is the prematurity of the baby, and operative delivery takes place. The diabetic women proceed to cesarean section. Women with pre-gestational diabetes were overweight and older. Diabetes occurs during this pre-gestational time and lasts longer in women. This

diabetes occurs in women after pregnancy.

Gestational diabetes

Gestational diabetes is detected for the first time during pregnancy. Gestational diabetes affects the glucose level of pregnant women as well as the body. It can also affect the baby's fitness. During the time of pregnancy, gestational diabetes can easily be controlled by eating healthy food, exercising, and taking proper medication. The symptoms of diabetes do not quickly happen, given some unnoticeable signs occur, like increased thirst and more frequent urination studied. Early and accurate diagnosis of diabetes mellitus, especially during its initial development, is challenging for medical professionals. Artificial intelligence and machine learning techniques, providing a reference, can help them gain preliminary knowledge about this disease and reduce their workload accordingly. Significant numbers of research have been performed to predict diabetes automatically using machine learning and ensemble techniques. Most of these works employed the open-source Pima Indian dataset.

Unsupervised Learning

To teach a model in unsupervised machine learning, it does not rely on labels or a predefined output variable. Instead, the program must identify data patterns and relationships. In an unsupervised learning setting, knowledge acquisition may be achieved by focusing on the Xs, representing the input data, and an overall performance assessment function. The dataset consists of an experimental collection of vectors with no associated functional values. Due to the lack of labeling in the given cases, the learner cannot check the correctness of the structure produced by the corresponding algorithm. According to Hofmann, supervised learning tasks require labels in data, which can be established with unsupervised learning methods. Again, under unsupervised learning, there are two categories of problems: association and clustering. Unsupervised learning algorithms also have several categories, some of which are k-means clustering and hierarchical clustering. The mechanism behind unsupervised learning is described below in the diagram.

RESEARCH METHODOLOGY

Analyse and design classification methods for data which involves less number of the distance computation process. The main focus is on increasing true positive rates and true negative rates for imbalanced datasets. To emphasis on improving the accuracy for standard datasets as well as for missing value datasets. This means that there is an ever-present and ever-increasing scope to improve the quality and accuracy of the analysis of this data. Though a number of classification algorithms are available to us none is perfect or even good enough to be able to simply handle the variety of data being produced. A lot of work is therefore ongoing in this field for improving these techniques on a quotidian basis. The unwanted existence of outliers or missing values in the training dataset often decreases the accuracy of a model. Poor performance when dealing with datasets with missing values. The requirement of finding distances from a query point to all data points, whenever the classification of a query (unknown data) point is required. Large amounts of data are created on a daily basis in many different forms, sizes and typologies. Although much work has been done on this concern, still there is a scope of improvement in performance with

respect to accuracy. To modify the existing classification algorithm for handling standard and large datasets with comparatively lesser requirement of the distance computation process. To transform the existing algorithm so that it can handle imbalanced datasets and detect outliers much more efficiently.

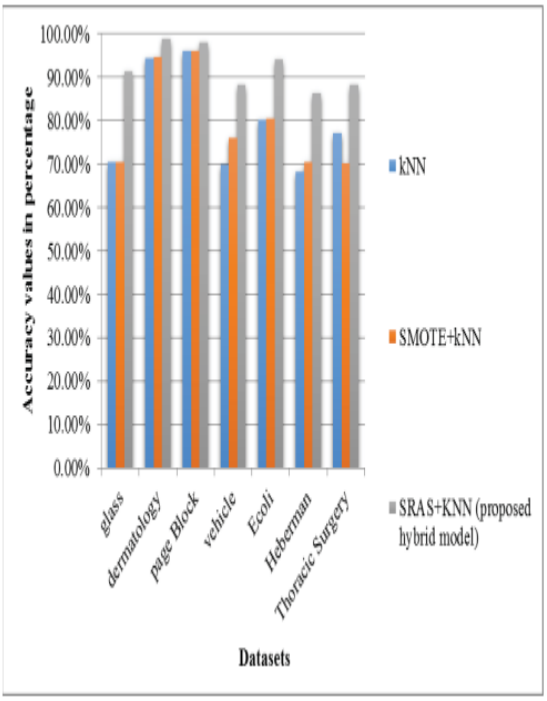
RESULTS AND DISCUSSIONS

The accuracy values of the proposed technique versus the two well-known existing techniques k NN, SMOTE + k NN, and the bold values represent the highest values obtained during the experiments done individually on each technique with respect to the datasets.

Table 1: Accuracy Values of the three Techniques

Data sets	*k* NN	SMOTE+ *k* NN	SRAS + *k* NN
Glass	70.56	70.40	**91.48**
Dermatology	94.54	94.56	**98.96**
Page Block	96.02	96.16	**98.15**
Vehicle	69.86	76.08	**88.23**
Ecoli	80.36	80.47	**94.08**
Haberman	68.3	70.54	**86.30**
Thoracic Surgery	77.23	70.19	**88.33**

Graph 1 is the corresponding graph of the accuracy values shown in Table 1. The grey bar in the graph shows the highest accuracy values which are obtained by the proposed technique SRAS + k NN.



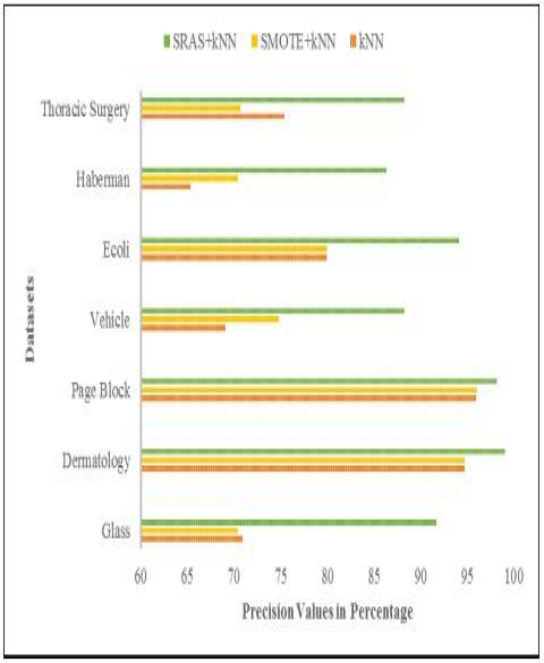
Graph 1: Accuracy of the three Techniques

Table 2 describes the classification performance values such as F-measure, precision, and recall obtained after executing the datasets on the three techniques in percentage, the green lines in the graphs show the higher values obtained. The bold values in all the tables show the highest values obtained. Higher values signify that the classification results are more accurate than the other two techniques.

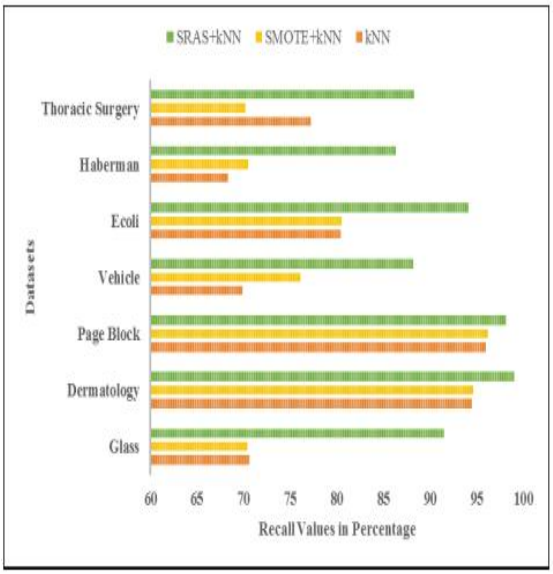
Table 2: Comparison of Precision, Recall and F measure Values of the three Techniques

	Precision			Recall			F measure		
Dataset	*k* NN	SMOTE+ *k* NN	SRAS + *k* NN	*k* NN	SMOTE + *k* NN	SRAS+ *k* NN	*k* NN	SMOTE + *k* NN	SRAS+ *k* NN

Glass	70.9	70.4	**91.7**	70.6	70.4	**91.5**	70.4	70.2	**91.5**
Dermatology	94.7	94.7	**99**	94.5	94.6	**99.0**	94.5	94.5	**99.0**
Page Block	95.9	96.0	**98.1**	96.0	96.2	**98.1**	95.9	96.1	**98.1**
Vehicle	69.1	74.8	**88.2**	69.9	76.1	**88.2**	69.4	75.4	**88.2**
Ecoli	79.9	79.9	**94.1**	80.4	80.5	**94.1**	80.1	80.1	**94.0**
Haberman	65.3	70.4	**86.3**	68.3	70.5	**86.3**	66.4	70.5	**86.3**
Thoracic Surgery	75.4	70.7	**88.2**	77.2	70.2	**88.3**	76.3	70.4	**88.2**



Graph 2: Precision Values of k NN, k NN and SMOTE and the Proposed Hybrid Technique SRAS with k NN



Graph 3: Recall Values of k NN, k NN and SMOTE and the Proposed Hybrid Technique SRAS with k NN

Describes the ROC curve area values of the three techniques. Graph 3 is the corresponding ROC area graph. The proposed technique outperforms the other two techniques significantly.

CONCLUSIONS

The term data mining in databases has been embraced for a field of research managing the programmed disclosure of verifiable data or information inside databases. Various fields, for example, promoting client relationship, designing, drug, fraud analysis, expert prediction, web mining use information mining. Databases are rich with unseen data, which could be utilized for intelligent decision making. Data mining refers to a framework that has the ability to consequently take in information as a matter of fact in different ways. A classification is a form of data exploration that are used to mine significant information or to estimate future data inclinations. The developed preprocessing techniques is good not only for k NN but for other classification models moreover. The working of each classification algorithm has been studied thoroughly and the results obtained during the experiment with regard to accuracy, recall, specificity, and precision have been recorded and further analysed. Imbalanced data is created in situations where the data is produced in many areas such as medical data, risk management like insurance, anomaly detection. This technique transforms the imbalanced dataset to a balanced dataset and discards the unwanted and irrelevant attributes from the balanced data. The quality of classification modelling can be greatly enhanced by identifying and excision of these values. So, a hybrid preprocessing technique using resample method and interquartile range method (IQR) which manages imbalanced data and outliers has been developed.

REFERENCES

1. *Mamta Bansal (2023), "Diabetes prediction model using data mining techniques", Measurement: Sensors, issn:2665-9174, vol.25, https://doi.org/10.1016/j.measen.2022.100605*
2. *Pijush Kanti Dutta Pramanik (2023), "An ensemble learning approach for diabetes prediction using boosting techniques", Front Genet., issn:1664-8021, vol.14, doi: 10.3389/fgene.2023.1252159*
3. *Worku Gachena Negera (2021), "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)", Diagnostics (Basel)., vol.11(9), pages.1714. doi: 10.3390/diagnostics11091714*
4. *Saad Althobaiti (2024), "An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making", Alexandria Engineering Journal, issn:2090-2670, vol.94, pages.311-324.https://doi.org/10.1016/j.aej.2024.03.044*
5. *Beschi Raja (2019), "Diabetics Prediction using Gradient Boosted Classifier", International Journal of Engineering and Advanced Technology (IJEAT), issn:2249-8958, vol.9, issue.1,*
6. *Silvia Aparicio Obregon (2024), "Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model", Scientific Reports, issn: 2045-2322, vol.14,https://doi.org/10.1038/s41598-024-74357-w*
7. *Arjinder Sethi (2015), "Acute Complications of Myocardial Infarction in the Current Era: Diagnosis and Management", Journal of Investigative Medicine, issn:1708-8267, vol.63, issue.7,https://doi.org/10.1097/JIM.0000000000000232*
8. *Intaek Kim (2021), "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm", Int. J. Environ. Res. Public Health, issn:1660-4601, vol.18(6), pages.3317. https://doi.org/10.3390/ijerph18063317*
9. *Md Maniruzzaman (2017), "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm", Comput Methods Programs*

Biomed., issn:0169-2607, vol.152, pages.23-24.doi: 10.1016/j.cmpb.2017.09.004

10. *Harman S Suri (2018), "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers", J Med Syst., issn:1573-689X, vol.42(5), pages.92.doi: 10.1007/s10916-018-0940-7*