2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10725693

A Comprehensive and Comparative Examination of Machine Learning Techniques for Diabetes Mellitus Prediction

1st Ajay Kumar
Department of Computer Science and Engineering
Manipal University Jaipur
Rajasthan 303007, India
ajay3789@gmail.com

2nd Anmol Singh Gill
Department of Computer Science and Engineering
Manipal University Jaipur
Rajasthan 303007, India
anmolsgill2550@gmail.com

3rd Jay Prakash Singh
Department of Computer Science and Engineering
Manipal University Jaipur
Rajasthan 303007, India
jaykiit.research@gmail.com

4th Debolina Ghosh
Department of Information Technology
Manipal University Jaipur
Rajasthan 303007, India
debolina442@gmail.com

***Abstract*—Selecting the optimal machine learning (ML) approach is paramount in healthcare and medical informatics to ensure precise predictions that align with the dataset's inherent characteristics. In this study, we explore the field of diabetes prognosis by employing five distinct ML models: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and k-Nearest Neighbors (k-NN), to evaluate their efficacy in predicting future diabetic diagnoses. Our findings reveal that the RF classifier emerges as the most effective model, achieving an impressive accuracy rate of 92.23%. This exceptional performance is attributed to its proficiency in navigating the intricate web of complex and non-linear relationships embedded within the input features. In contrast, the LR model demonstrated less favorable outcomes, with a comparatively modest accuracy score of 74.42%. These results underscore the importance of utilizing sophisticated ML algorithms capable of discerning nuanced patterns in healthcare datasets, thereby equipping clinicians with robust predictive tools to enhance patient care and prognosis.**

***Index Terms*—Machine Learning, Dataset, Diabetes Mellitus, Prediction, Classifiers, Accuracy**

I. Introduction

Diabetes mellitus is an instability highlighted by the body's abnormally high blood sugar (glucose) levels as a outcome of diminished insulin production or utilization. People who urinate more frequently and have increased thirst may lose weight. Diabetes may cause nerve damage and a reduction in tactile perception. It also increases the risk of developing illnesses like chronic kidney disease, heart attacks, strokes, and visual impairment. Blood sugar tests are used by doctors to diagnose diabetes. Diets low in processed foods, refined carbohydrates (especially sugar), and saturated fats are recommended for people with diabetes. Generally speaking, blood sugar maintenance requires regular exercise, weight control, and medication. A higher-than-normal blood glucose level is indicative of diabetes mellitus. Towards distinguish it from diabetes insipidus, doctors often refer to this condition as diabetes mellitus rather than just "diabetes." Diabetes insipidus is a disorder characterized by urine production without affecting blood glucose levels [1]. Medical practitioners find it challenging to diagnose diabetes mellitus accurately and early, especially in the disease's early stages. With the use of artificial intelligence (AI) and machine learning (ML) techniques, they can gain insights into this illness and reduce their workload [2]–[6]. Many studies have been conducted to utilize ensemble and ML techniques for automatically predicting diabetes. Several of such initiatives used the open-source Pima Indian dataset. In a dataset with 800 records and 10 attributes, for instance, Mujumdar and Vaidehi [7] used ML algorithms, and logistic regression (LR) obtained the best accuracy of 96%. Moreover, the accuracy of ML algorithms varies depending on the dataset, as evidenced by their utilization of two distinct datasets resulting in differing accuracies. Bhat et al., [8] explored an early prediction of diabetes using ML, applying six such as random forest (RF), multi-layer perceptron (MLP), support vector machine (SVM), gradient boost (GB), decision tree (DT), and LR algorithms to the PIMA dataset, which resulted in a 98% accuracy for RF. Hassan et al., [9] utilized the Pima Indian dataset and ensemble method-based ML techniques to predict diabetes, achieving an area under the curve (AUC) value of 0.95. Jackins et al., [10] divided the sample into four classes and applied three (naive bayes, SVM and Light-GB) algorithms on a dataset of 520 patients at Sylhet Diabetes Hospital, Bangladesh, with

979-8-3503-6066-0/24/$31.00 ©2024 IEEE
15th ICCCNT IEEE Conference,
June 24-28, 2024,
IIT - Mandi, Kamand, India

SVM achieving the highest accuracy of 96.54%. Sneha and Gangil, [12] conducted ML algorithms on a dataset from the UCI machine repository. Their study explored enhancing model accuracy by calculating attribute correlations and excluding highly correlated ones. This analysis of relevant literature suggests the successful integration of diverse ML algorithms and preprocessing methods for automated diabetes identification. Most studies utilized the publicly-available Pima Indian dataset, focusing solely on accuracy without enhancing predictability. Hence, we evaluate our prediction system using metrics including accuracy, precision, recall, and F1 score. Various ML-based classification algorithms like LR, RF, SVM, DT, and KNN are employed. The model with the highest performance is selected. The paper's structure includes discussions supported by flowcharts and figures for the suggested diabetes prediction system in Section 2, while Sections 3 and 4 present conclusive survey findings.

II. PROPOSED SYSTEM

This section elucidates the design process of the proposed automatic diabetes prediction system and the implementation of various ML techniques. The accompanying Fig.1, illustrates the different stages of the survey. Initially, data collection is conducted, followed by preprocessing to render the data suitable for ML algorithms. This preprocessing entails handling missing and duplicated values, as well as feature scaling to prevent the dominance of one feature over others. Additionally, the correlation between attributes is assessed and managed to mitigate overfitting. The updated dataset is then partitioned into training and test sets. Subsequently, diverse classification algorithms are applied to the training data, and the model yielding the highest accuracy on the test set is selected. This iterative process ensures robust model selection for diabetes prediction. Furthermore, the workflow incorporates rigorous data preparation and model evaluation to enhance prediction accuracy. By systematically addressing data quality issues and selecting appropriate ML algorithms, the system aims to achieve reliable diabetes prediction outcomes. The division of the dataset into training and test subsets facilitates unbiased model evaluation, enabling the identification of the most effective classifier.

A. Dataset

The diabetes prediction dataset [20], serves as the cornerstone of this paper's research endeavors, acting as the primary source of data. The PIMA Indians Diabetes Dataset, commonly used for benchmarking ML models in medical research, includes 768 instances with 8 input features (such as pregnancies, glucose levels, and BMI) and 1 output feature indicating diabetes presence. Also, the Pima Indians Diabetes Database directs readers to further details about this dataset in the corresponding reference, summarised in Table I. Emphasizing its pivotal role, the dataset's utilization underscores its relevance to the paper's objectives and findings. Within this dataset lies a comprehensive array of meticulously gathered health-related characteristics, highlighting its depth and breadth. These characteristics form the foundation for developing predictive models aimed at forecasting or identifying diabetes-related outcomes. By leveraging the dataset's diverse set of health-related variables, these models hold promise in pinpointing individuals at risk of diabetes, thereby offering valuable insights into preventative measures and early intervention strategies. Overall, the dataset's richness and utility underscore its significance in advancing our understanding and management of diabetes, as shown in Fig.,2.

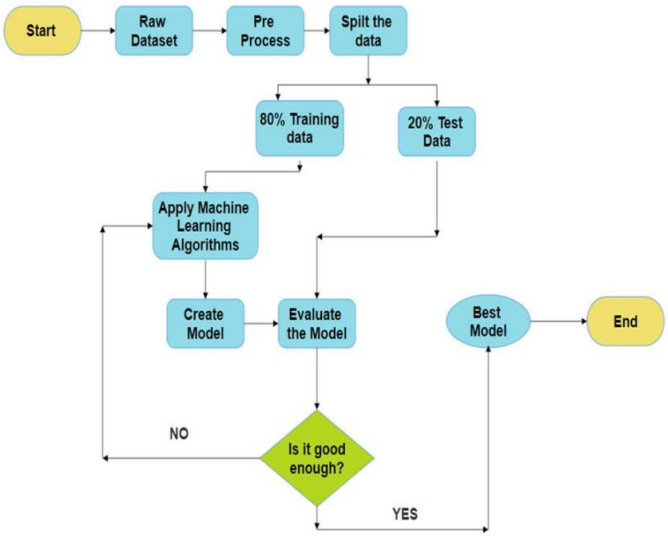

Fig. 1. Flowchart of the proposed system.

TABLE I
DESCRIPTION OF PIMA DATASET

Sl.No	Features	Summary
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test
3	Skin thickness	Triceps skin fold thickness (mm)
4	Blood pressure	Diastolic blood pressure (mm Hg)
5	Insulin	2-Hour serum insulin
6	BMI	Body mass index
7	Diabetes pedigree function	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	1 if diabetes present; else 0

B. Exploratory data analysis (EDA)

Behrens [21] explored the concept of employing an EDA, frequently utilizing data visualization techniques, to scrutinize, analyze, and summarize key characteristics of datasets, as depicted in Fig. 3.

- *Understanding the variables*: A closer look at catalogues, field descriptions, and metadata in greater detail can help identify incomplete or missing data and provide insight into what each field represents. There are no null and duplicate values in this dataset.
- *Detecting outliers*: Early detection of outliers is vital as they can distort dataset analysis. Common techniques in-

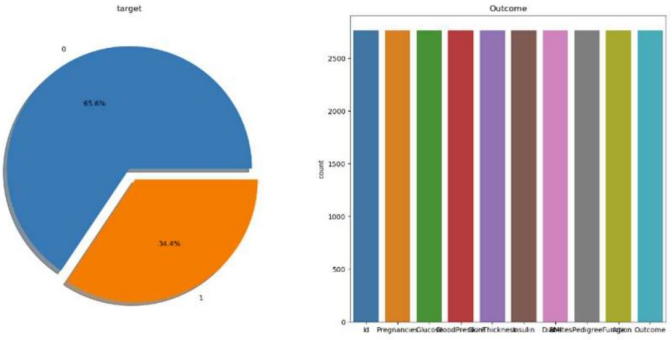

Fig. 2. Pie chart of the target variable.

	count	mean	std	min	25%	50%	75%	max
Id	2768.0	1384.500000	799.197097	1.000	692.750	1384.500	2076.250	2768.00
Pregnancies	2768.0	3.742775	3.323801	0.000	1.000	3.000	6.000	17.00
Glucose	2768.0	121.102601	32.036508	0.000	99.000	117.000	141.000	199.00
BloodPressure	2768.0	69.134393	19.231438	0.000	62.000	72.000	80.000	122.00
SkinThickness	2768.0	20.824422	16.059596	0.000	0.000	23.000	32.000	110.00
Insulin	2768.0	80.127890	112.301933	0.000	0.000	37.000	130.000	846.00
BMI	2768.0	32.137392	8.076127	0.000	27.300	32.200	36.625	80.60
DiabetesPedigreeFunction	2768.0	0.471193	0.325669	0.078	0.244	0.375	0.624	2.42
Age	2768.0	33.132225	11.777230	21.000	24.000	29.000	40.000	81.00
Outcome	2768.0	0.343931	0.475104	0.000	0.000	0.000	1.000	1.00

Fig. 3. Invaluable insights into the characteristics of EDA

clude data visualization, numerical methods, interquartile ranges, and hypothesis testing. Boxplots, as shown in Fig.4, aid outlier identification.

- *Analyze patterns and relationships*: Plotting a dataset in various ways simplifies the process of finding and examining patterns and relationships between variables, as shown in Fig. 5.
- *Correlation*: An insight-based statistical approach indicating that one variable moves or changes with another is correlation, offering insight into the strength of their relationship. Fig. 6, illustrates feature-outcome correspondence.

C. Data pre-processing

In ML, data pre-processing involves cleaning and organizing raw data to prepare it for building and training models [18]. The dataset has been cleansed of null, duplicate, and outlier data. Additionally, the attribute ”Id” has been removed due to its lack of correlation with the outcome variable.

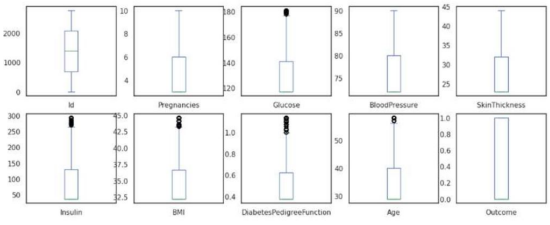

Fig. 4. Box-Plots after removing Outliers

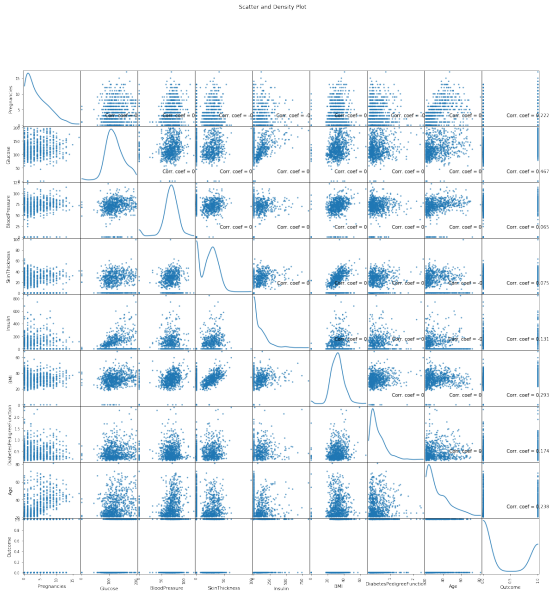

Fig. 5. Visualizing data variability with scatterplots and density plots

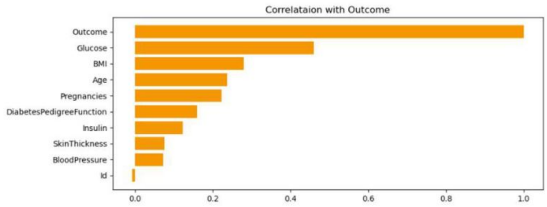

Fig. 6. Visualizing of correlation with outcome

III. MACHINE LEARNING CLASSIFIERS

ML classifiers are algorithms used to categorize data based on patterns and features, aiding in automated decision-making processes.

A. Logistic Regression

LR utilizes a dataset of independent variables to estimate the probability of an event occurrence, such as the presence or absence of diabetes. The resulting probability, ranging from 0 to 1, is derived from a logit transformation applied to the odds, representing the ratio of success probability to failure probability in LR analysis.

B. Support Vector Machine

SVM is a supervised learning technique for discriminative classification. It aims to establish an optimal decision boundary, or hyperplane, to categorize future data points efficiently. This hyperplane separates n-dimensional space into distinct classes. By selecting extreme vectors, known as support vectors, SVM constructs this hyperplane, facilitating accurate classification.

C. Decision Tree

Although DT is commonly associated with solving classification problems, it can also address regression problems. Fig.

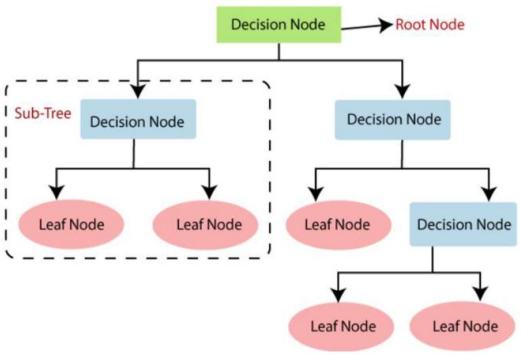

Fig. 7. Visualizing of decision tree

7, illustrates the tree-structured classifier utilizes internal nodes to represent dataset features, branches for decision rules, and leaf nodes for outcomes. Decision nodes facilitate decision-making with multiple branches, while leaf nodes signify final decisions without additional branches. The dataset's characteristics inform these decisions or tests within the DT structure.

D. Random Forrest Classifier

RF is a supervised learning method capable of performing both classification and regression tasks. It employs a bagging technique to create random samples of features. Unlike DT, which follows a deterministic approach, RF randomly selects features for identifying root nodes and dividing feature nodes, distinguishing it from traditional DT algorithms.

E. k-NN

k-NN is a simple yet widely used supervised ML algorithm employed in missing value imputation, classification, and regression tasks. It operates on the principle that observations nearest to a given data point are the most similar, enabling the categorization of unseen points based on the values of nearby observations. The user determines the number of close observations considered by choosing K(=5).

IV. RESULTS AND DISCUSSION

This section presents the findings and discussions of the suggested automated diabetes prediction system. Also, 10-fold cross-validation, which involves splitting the dataset into 10 parts and training the model 10 times, is a popular technique for evaluating classification models for predicting diabetes. This robust assessment reduces overfitting and ensures reliable performance metrics such as accuracy, precision, recall, and F1-score, which are crucial for medical informatics. Additionally, the confusion matrix evaluates a classification model's performance by detailing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics provide a comprehensive assessment of the model's predictive ability. The formulas for these measurements are provided as part of the analysis in Equation (1)-(2)-(3).

TABLE II
PERFORMANCE METRICS OF VARIOUS CLASSIFIERS

Classifier	Precision	Recall	F1 Score	Accuracy (%)
Logistic regression	0.67	0.44	0.53	74.62
KNN	0.72	0.68	0.70	81.18
Random forest	0.92	0.83	0.87	92.23
Decision tree	0.75	0.57	0.65	79.87
SVM	0.70	0.43	0.53	75.38

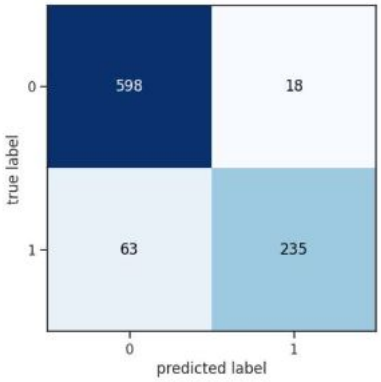

Fig. 8. Confusion matrix of RF model demonstrating robust performance

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - score = \frac{2RecallPrecision}{Recall + Precision} \tag{3}$$

In this context, TP signifies a positive prediction from the model aligning with a positive outcome. Conversely, FP denotes a positive prediction from the model when the outcome is negative. TN indicates both a negative outcome and a negative prediction from the model. FN represents a positive outcome, despite the model predicting a negative outcome. Throughout this study, all machine learning models have been evaluated using the holdout validation approach, employing a stratified 8:2 train-test split. Table II presents a comparison of various performance metrics of our classifiers on the dataset. Based on this table and its remarks about the confusion matrix and ROC curve shown in Fig.8 -9, the random forest classifier demonstrated the highest overall performance, achieving 92% accuracy, and F1 scores and recall values of 0.87 and 0.83, respectively. It is shown in the performance metrics of different models in Fig. 10. Moreover the objective of this research is to utilize ML techniques for automated prediction of diabetes mellitus onset.

V. CONCLUSION

Diabetes significantly impacts both life expectancy and quality, making early prediction crucial for reducing associated risks and complications. This study presents an automated

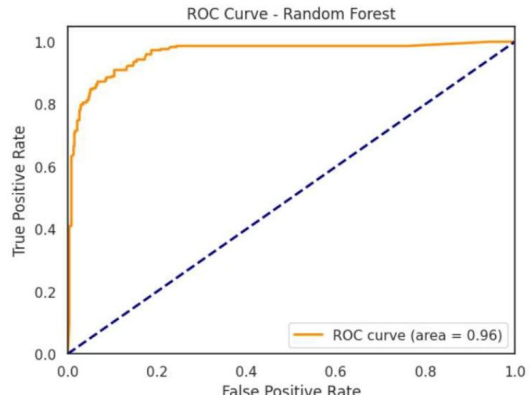

Fig. 9. ROC curve of RF model demonstrating robust performance

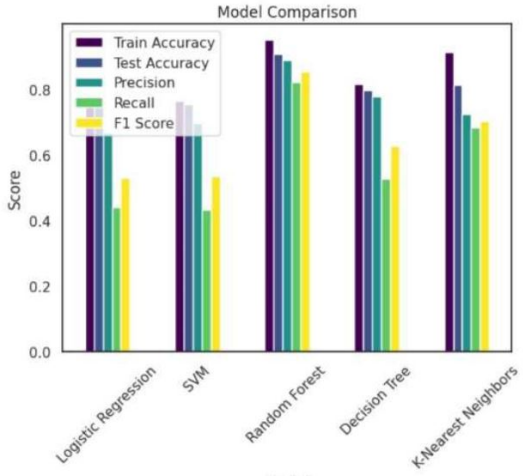

Fig. 10. Visualisation of performance metrics of different models

diabetes prediction system employing various machine learning approaches. Five ML classifiers are applied, analyzed, and tested on different parameters, with the RF Classification Algorithm achieving the highest accuracy of 92%. This model, along with other ML techniques, holds promise for identifying or diagnosing various diseases. Future research could expand upon and enhance this work for diabetes research and explore comparisons between ML algorithms and other predictive technologies, such as clinical decision support systems, for diabetes diagnosis and management.

ACKNOWLEDGMENT

The authors sincerely thank to Department of Computer Science and Engineering, Manipal University Jaipur, Rajasthan, India, for facilitating the computing facility to execute the work

REFERENCES

[1] National Heart, Institute, B. & Others National Institute of Diabetes and Digestive and Kidney Diseases. *Clinical Guidelines On The Identification, Evaluation And Treatment Of Overweight And Obesity In Adults. The Evidence Report. Bethesda: National Institutes Of Health*. **1** pp. 228 (1998)

[2] Kumar, A. & Gorai, A. Application of transfer learning of deep CNN model for classification of time-series satellite images to assess the long-term impacts of coal mining activities on land-use patterns. *Geocarto International*. **37**, 11420-11440 (2022)

[3] Kumar, A. & Gorai, A. Design of an optimized deep learning algorithm for automatic classification of high-resolution satellite dataset (LISS IV) for studying land-use patterns in a mining region. *Computers Geosciences*. **170** pp. 105251 (2023)

[4] Kumar, A. & Gorai, A. A comparative evaluation of deep convolutional neural network and deep neural network-based land use/land cover classifications of mining regions using fused multi-sensor satellite data. *Advances In Space Research*. **72**, 4663-4676 (2023)

[5] Kumar, A. & Gorai, A. Development of a deep convolutional neural network model for detection and delineation of coal mining regions. *Earth Science Informatics*. **16**, 1151-1171 (2023)

[6] Kumar, A. A Promising Automatic System for studying of Coal Mine Surfaces using Sentinel-2 Data to Assess a Classification on a Pixel-based Pattern. *Journal Of Mining And Environment*. **15**, 41-54 (2024)

[7] Mujumdar, A. & Vaidehi, V. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*. **165** pp. 292-299 (2019)

[8] Bhat, S., Selvam, V., Ansari, G., Ansari, M., Rahman, M. & Others Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora. *Computational Intelligence And Neuroscience*. **2022** (2022)

[9] Hasan, M., Alam, M., Das, D., Hossain, E. & Hasan, M. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*. **8** pp. 76516-76531 (2020)

[10] Jackins, V., Vimal, S., Kaliappan, M. & Lee, M. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal Of Supercomputing*. **77**, 5198-5219 (2021)

[11] Xue, J., Min, F. & Ma, F. Research on diabetes prediction method based on machine learning. *Journal Of Physics: Conference Series*. **1684**, 012062 (2020)

[12] Sneha, N. & Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal Of Big Data*. **6**, 1-19 (2019)

[13] Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F. & Ren, Z. Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal Of Environmental Research And Public Health*. **19**, 15027 (2022)

[14] Paliwal, M. & Saraswat, P. Research on Diabetes Prediction Method Based on Machine Learning. *2022 2nd International Conference On Technological Advancements In Computational Sciences (ICTACS)*. pp. 415-419 (2022)

[15] Tasin, I., Nabil, T., Islam, S. & Khan, R. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*. **10**, 1-10 (2023)

[16] Prabhu, P. & Selvabharathi, S. Deep belief neural network model for prediction of diabetes mellitus. *2019 3rd International Conference On Imaging, Signal Processing And Communication (ICISPC)*. pp. 138-142 (2019)

[17] Negi, A. & Jaiswal, V. A first attempt to develop a diabetes prediction method based on different global datasets. *2016 Fourth International Conference On Parallel, Distributed And Grid Computing (PDGC)*. pp. 237-241 (2016)

[18] Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G. & Wunsch, D. Computational learning approaches to data analytics in biomedical applications. (Academic Press,2019)

[19] Kumar, S. & Chong, I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *International Journal Of Environmental Research And Public Health*. **15**, 2907 (2018)

[20] Smith, J., Everhart, J., Dickson, W., Knowler, W. & Johannes, R. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings Of The Annual Symposium On Computer Application In Medical Care*. pp. 261 (1988)

[21] Behrens, J. Principles and procedures of exploratory data analysis.. *Psychological Methods*. **2**, 131 (1997)