Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Research on Integrating Artificial Intelligence and Data Preprocessing for Diabetes Prediction

Bhawana Saraswat[1], Dr. Gayathri R[2], Prabhat Sharma[3], Dr Kumud Saxena[4], Sridisha Banerjee[5], Hitesh Kalra[6]

[1]Scholar, Department of Computer Science & Engineering, Sanskriti University, Mathura, Uttar Pradesh, India, Email Id- bhawnasaraswatphd@sanskriti.edu.in
[2]Assistant Professor, Department of OB & HRM, JAIN (Deemed-to-be Univesity), Bangalore, Karnataka, India, Email Id- dr.gayathri_r@cms.ac.in
[3]Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India, Email Id- prabhat.sharma.orp@chitkara.edu.in
[4]Professor and HOD, Department of Computer Science, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, Email Id- kumud.saxena@niet.co.in
[5]Assistant Professor, Department of Management Studies, Vivekananda Global University, Jaipur, India, Email Id- sridisha.banerjee@vgu.ac.in
[6]Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh-174103 India, Email Id- hitesh.kalra.orp@chitkara.edu.in

Introduction: Hyperglycemia, a metabolic disorder brought on by the body's incapacity to make and respond to insulin, is the hallmark of diabetes mellitus.
Methods: This study focuses on employing Artificial Intelligence (AI) approaches combination with improved data preparation methods to improve predicted accuracy of diabetes diagnosis utilizing the PIMA Indians Diabetic database. The preprocessing step comprises identifying and discarding uncommon incidents within the dataset, assuring the elimination of outliers that might negatively affect model performance. Missing value imputation methods are used to manage incomplete data, adopting methodologies such as data-driven imputation to improve the dataset's completeness. The proposed Extreme Gradient Boosting-K-Nearest Neighbor (XGBoost-KNN) technique was chosen for its ability to handle complicated connections within data as well as whilst KNN was chosen for its ability to capture local patterns.
Result: The findings provide a comparison between the efficacy of the suggested model, XGBoost-KNN, and that of conventional Machine Learning (ML) methods such as Logistic-Regression (LR), Support-Vector-Machines (SVM), and Naïve-Bayes (NB), making use of relevant performance measures such as 'accuracy', 'precision', 'recall', and 'F1-score'. The proposed XGBoost-KNN model shows good results, suggesting its potential as an accurate and reliable method for detecting persons at risk of acquiring diabetes.
Conclusion: These results have larger significance in the field of healthcare, providing the foundation for the development of preemptive treatments and individualized healthcare methods to reduce effect of diabetes on general population. The findings show that BRF-MOANN outperforms

traditional methods to provide more comprehensive and precise diagnosis of severity.

Keywords: Artificial Intelligence (AI), Diabetes mellitus, Machine Learning (ML), Extreme Gradient Boosting-K-Nearest Neighbor (XGBoost-KNN).

1. Introduction

Diabetes mellitus is a global disease that results from uncontrolled glucose levels in the blood and excessive urine production due to insulin deficiency or improper consumption. [(1)] High blood glucose levels are a defining feature of diabetes, a chronic illness that impairs the kidneys, eyes, and heart. Type 1 diabetes also known as juvenile diabetes, is insulin-dependent, whereas type 2 diabetes is insulin resistance. [(2)] Predictive modeling and Artificial Intelligence (AI) can enhance care in various clinical areas, such as diagnosis, risk assessment, lifestyle management, and home monitoring. Effective AI-driven techniques to improve care could impact common chronic diseases like diabetes mellitus, which have high morbidity and death. [(3)] AI aims at generating complex assumptions on an extensive quantity of data. [(4)] Big databases are seen in the healthcare industry. Data from unstructured, partially organized, or well-organized databases may be present. Big data mining examines massive data sets and finds unrecognized patterns and facts for the purpose to derive knowledge based on the information that was provided. [(5)] Insulin production is impacted by diabetes; an inflammatory illness by obesity and elevated blood glucose levels. The World Health Organization estimates that 422 million people globally, primarily in low-income countries, suffer from Type 2 diabetes. Although early detection can save lives, the prevalence is rising worldwide. [(6)] Diabetes that is not controlled periodically leads to a state of hyper or elevated glucose levels, which damages many organs and tissues over time, including neurons and arterial capillaries. [(7)] Different types of data are gathered and saved by the Information Input layer from other places, including Electronic Health Records (EHR), in a format that can be used as input for machine learning algorithms. [(8)]

The work [(9)] investigated the creation of models by utilizing various machine learning categorization methods, can extremely accurately anticipate the presence of insulin resistance in individuals. Genetic algorithms are proposed to provide the best possible results for classifying disease risk. The article [(10)] analyzed the current state of the art in data mining-based identification of diabetes and estimation, looks into diabetes management options, and offers a categorization and evaluation of used methods. The study [(11)] examined the data mining methods used to forecast the likelihood of diabetes. 520 diabetes patient cases were examined using the SVM, NB, and LR, gets closer. The most accurate method was RF. The study [(12)] employed feature selection using individual and ensemble approaches, utilizing LR and the Python IDE on two important datasets: PIMA Indians Diabetes and Vanderbilt.

The goal of this study is to increase the expected success rate for diabetes diagnosis using the PIMA Indians Diabetics database by combining Artificial Intelligence (AI) techniques with advances in data preparation techniques.

2. Methodology

The gathered data used to forecast, evaluate, and treat diseases effectively. Categorization models handle value-based challenges see Figure 1. The PIMA dataset is used in research to build Machine Learning (ML) architecture for data discovery, measurements, and technology employed to assess the architecture.

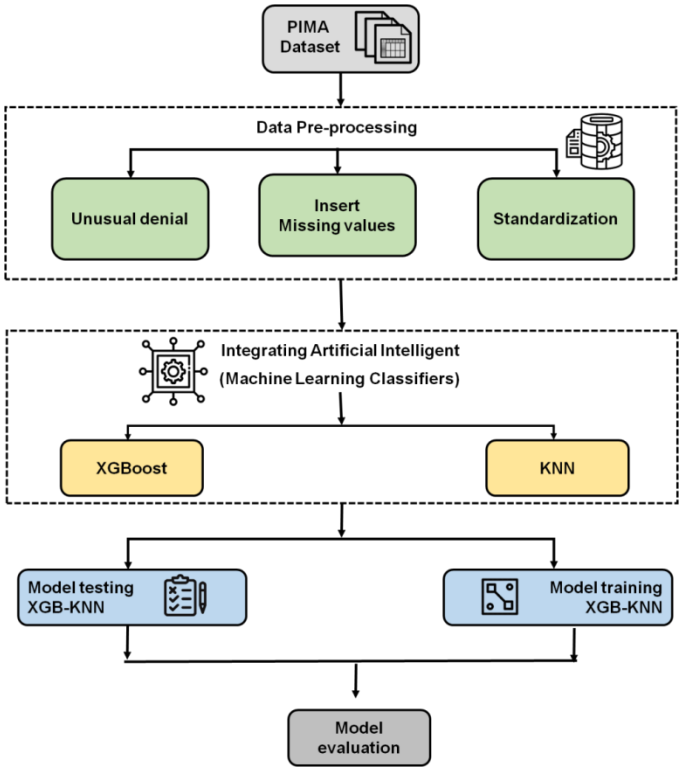

Figure 1. Work Flow Model [Source: Author]

PIMA Dataset

The PIMA Indians Diabetic dataset, which comprises 512 female diabetic patients from the Pima Indian community, was used to construct and evaluate the machine learning models [(13)]. Concerning eight features, the current set shall consist of 276 patients with diabetes and 236 without diabetic patients. For a synopsis of the statistics and interpretations of the attributes, see Table 1 and Figure 2. The Pedigree Factor was estimated per (1).

$$\text{Pedigree} = \frac{\sum_p R_p(88-\text{ADM}_p)+20}{\sum_q R_q(\text{ALC}_q-14)+50} \quad (1)$$

Where p and q represent relatives with and without diabetes, R represents the percentage of shared genes among relatives: 0,500 for parents or full siblings, 0,250 for half-siblings, grandparents, uncles, and 0,125 for half-aunts, half-uncles, or first cousins, ADM_p and ACL_q represent relatives' ages in years at diagnosis and last non-diabetic test.

Table 1. Diabetic Patient Group Analysis [Source: Author]

Attributes	Mean± Std
Pressure	0,98 ± 0,96
Insulin	79,36 ± 119,34
Body Mass Index	38,00 ± 7,82
Glucose	140,00 ± 39,64
Age	35,76 ± 13,43

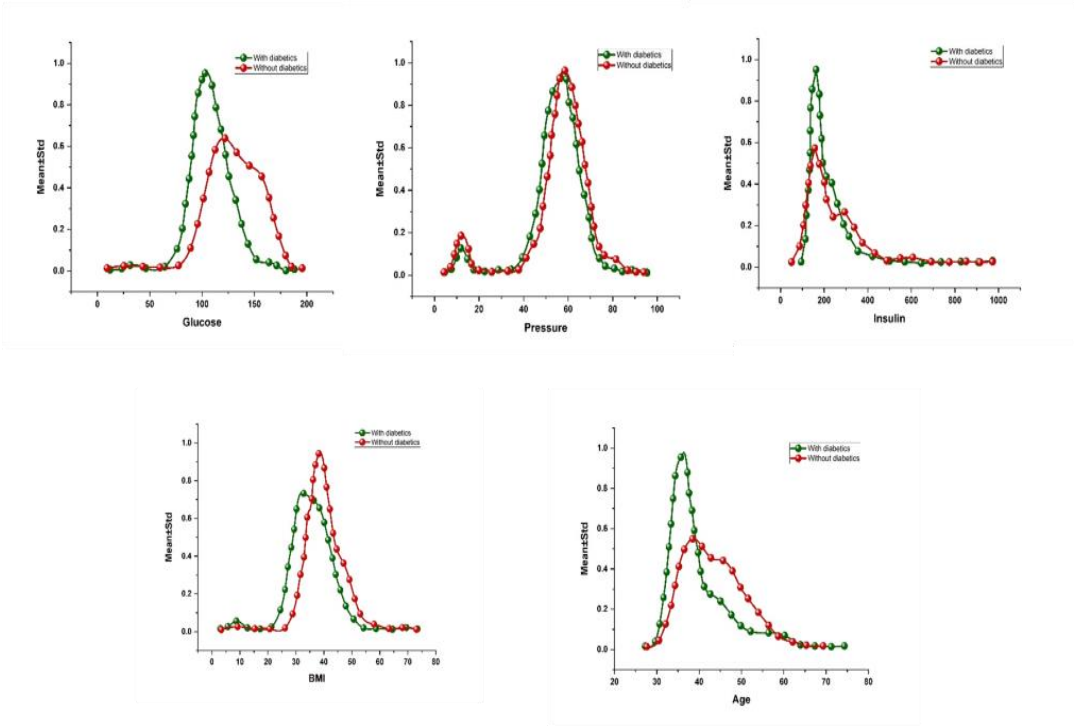

Figure 2. Statistical Analysis of Diabetes Prediction (Green- With Diabetes Red-Without Diabetes) [Source: Author]

Data Preprocessing

The suggested approach relies on preparing raw data, as data quality directly influences classifier learning. In the proposed system, preprocessing involves unusual events denial (B), insert missing value (O), standardization (K), and characteristic decision-making.

Unusual Events Denial

A strikingly different observation is an unusual value. It must be excluded from data distribution since classifiers are sensitive to attribute range and distribution. In this literature, remarkable events rejection is mathematically expressed as (2).

$$B(y) = \begin{cases} y, \text{if} O_1 - 1.5 \times IQR \leq y \leq O_3 + 1.5 \times IQR \\ \text{reject, otherwise} \end{cases} \quad (2)$$

Where, y represents the characteristic matrix elements in m-dimensional spacey $\in R^m$. $O_1, O_3, \text{and} IQR$ Represent the first, third, and medians of characteristics, respectively,

where $y \in R^m . O_1, O_3, \text{and} IQR$

Insert Missing Value

After unusual values denial, features were processed to add missing values, which could cause inaccurate classifier predictions. In the suggested framework, missing or null values were imputed by attribute mean values instead of dropping them, as shown in (3). Blaming continuous data with the mean is helpful as it eliminates unusual values.

$$O(y)\begin{cases}\text{mean}(y), \text{if } y = \frac{\text{null}}{\text{missed}} \\ y, \text{otherwise}\end{cases} \quad (3)$$

Standardization

The standardization or Z-score equalization technique rescales describes achieving a normal distribution with zero mean and unit variance. The standardization (K) described in (4) also decreases data distribution skewness

$$K(y) = \frac{y-\bar{y}}{\sigma} \quad (4)$$

If yis the m-dimensional feature vector instance, theny $\in R^m$. The average and variance of the qualities are $y \in R^m . \bar{y} \in R^m \sigma \in K^m$. Feature standardization is unlikely to guarantee substantial gains in many ML models, such as tree-based models.Classifier accuracy rises with feature dimension. The performance of classifiers decreases when feature dimensions rise without increasing the number of samples. The term scour of complexity in ML refers to this situation. The multidimensional issue leads to sparser feature space, over fitting, and loss of generalization in classifiers.

Extreme Gradient Boosting (XGBoost)-K-Nearest Neighbors Using Diabetics Prediction

Diabetic prediction using training data is proposed by supervised classification learning is XG Boost-KNN. ML algorithms can create diabetes risk prediction models. These models can assess risk factors and deliver individualized assessments.

K-nearest neighbors (KNN)

The classification is supervised ML techniques. KNNs are used to categorize input data into pre-defined classes. KNN calculates the Euclidean distance function between pre-defined styles and each sample. After that, KNN selects the minimum nearest neighbors for each category. Using the nearest k neighbors, models are categorized. There are several sample distance functions. This paper uses a Euclidean distance Equation (5) most often.

$$t = \sqrt{\sum_{r=1}^{m}(Y_{1r} - Y_{2r})^2} \quad (5)$$

Where r is the total quantity of elements for each array, and Y_1 and Y2are input specimens. The dataset is divided, employing six KNNs. The following information is provided. The Fine KNN uses one neighbor to discriminate sample data, while the Medium KNN uses several neighbors. The algorithm will have low distinctiveness with this kind. With more neighbors than the Medium KNN, the Coarse KNN has the most distinguishing feature among the three varieties. Equation (6)'s Cosine proximity statistic is used in the Cosine

KNN. Equation (7) applies to the Cubic KNN using a cubic relationship gauge. The weight KNN utilizes Equation (8) proximity factor. The following three categories have the same number of neighbors as Medium KNN.

$$t = (1 - \frac{y_1 y_2'}{\sqrt{(y_1 y_1')(y_2 y_2')}}) \quad (6)$$

$$t = \sqrt[3]{\sum_{r=1}^{m} |Y_{1r} - Y_{2r}|^3} \quad (7)$$

$$t = \sqrt{\sum_{r=1}^{m} u_j (Y_{1r} - Y_{2r})^2} \quad (8)$$

Classifier accuracy increases with fewer neighbors. Although this could boost classifier structure, more than accurate classification of out-of-samples is needed.

EXtreme Gradient Boosting (XGB)

XGB is a reliable distributed machine learning platform that deploys the Gradient Boosted Trees method for improving tree-boosting algorithms. Decentralized configurations with a fast parallel tree layout result in a well-configured, fault-tolerant classifier. Simplifying further, it can process billions of autonomous programming models and many millions of samples with one node in Equation (9).

$$x = \sum_{j=1}^{m} (u_j . y_j) \quad (9)$$

XGBoost's ability to handle complex data and capture nonlinear correlations makes it useful for diabetes prediction. Data must be preprocessed, model hyper parameters tuned and model performance evaluated to provide precise and dependable predictions.

3. Result and Discussion

The evaluation matrix, efficiency, and training time for all XG Boost-KNN types with 276 input samples. For accuracy, Weighted, Medium platform Cubic, Cosine and Coarse KNN are listed by performance. KNN and XG boost are the latest Python algorithms in diabetic prediction models. The latest stable version, Python 3.x offers the latest features, improvements, and security updates, ensuring compatibility with the latest libraries and frameworks. Table 2 shows the outcomes diabetic prediction's result of the XGB-KNN With the highest evaluation matrix. The proposed new technique, XGB-KNN, outscored the existing techniques, LR[13], SVM[14] and NB[15].

Table 2. Result for XGB-KNN and Existing Models [Source: Author]

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	74,90	63,99	61,94	62,73
SVM	75,98	81,08	84,92	81,99
NB	74,12	60,99	64,98	63,45
XGB-KNN	94,87	98,00	99,04	96,85

The XGB-KNN model has a higher accuracy rate of 94,87 %, while SVM shows good performance in precision and recall. The choice of model depends on the task's specific goals and requirements. However, LR and NB models have lower overall performance. The task goals and constraints determine the model and precision and recall (F1 Score) may be

important, while accuracy may be sufficient. The easiest measure to understand is accuracy, the ratio of all correct forecasts to all predictions. It provides a general sense of the model's forecast accuracy rate. The (A) accuracy Equation (10) is:

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+TN+FP+FN}} \quad (10)$$

A statistic called (B) precision shows what percentages of positive determinations were truly accurate. Stated differently, it assesses the model's capacity to classify an example as negative even though it is positive. The precision is computed as follows in Equation (11):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (11)$$

The particular objectives and demands of the activity determine which model is best, with memory and precision playing a key role in certain situations. One important indicator of a model's performance is accuracy, which is defined as the proportion of real positive projections among all anticipated positives.

In comparison with existing techniques LR (63,99 %), SVM (81,08 %), and NB (60,99 %), the proposed approach, XGB-KNN, achieved the maximum precision of 98,00 %. Examine figure 3.

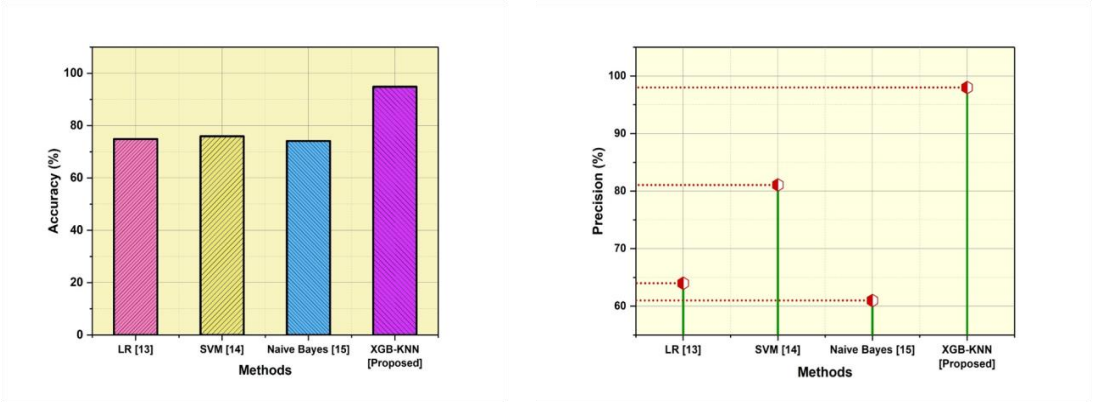

Figure 3. Outcomes of (A) Accuracy and (B) Precision [Source: Author]

The ratio of true positive forecasts to real positives, expressed as a percentage. Low rate of false negatives is indicative of high recall. A measure of recall indicates how much of the actual discoveries the algorithm has found. It shows that the model can find a dataset's pertinent examples. A recall is computed as follows in Equation (12):

$$\text{Recall} = \frac{\text{TP}}{\text{FN+TP}} \quad (12)$$

The suggested methodology, XGB-KNN, achieved the highest recall of 99,04 % when compared to the currently utilized methodologies, LR (61,94 %), SVM (84,92 %), and NB (64,98 %). See figure 4.

The geometric average of the recall and precision is called F1 score in Equation (13).

$$\text{F1 score} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \quad (13)$$

Figure 4 shows the Comparison of the evaluation matrix of the (A) recall and (B) F1 Score of existing and proposed approaches. The maximum F score of 96,85 % was attained by the XGB-KNN methodology, which is compared to the currently employed methodologies, LR (62,73 %), SVM (81,99 %), and NB (63,45 %).

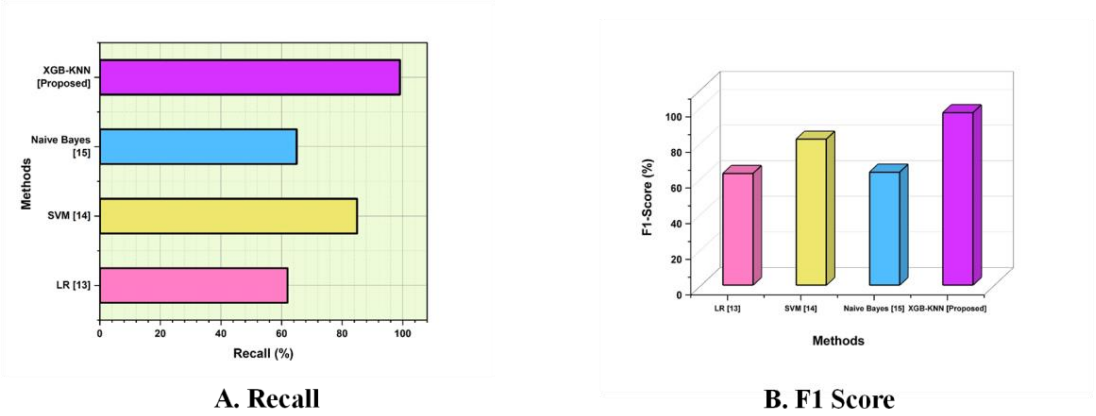

Figure 4. Comparison of (A) Recall and (B) F1 Score [Source: Author]

4. Discussion

To evaluate diabetes prediction, evaluation techniques and achievement criteria like recall, f-1 score, precision, and accuracy were employed. 80 percent of the 90 testing and 186 training observations were used for testing and 20 percent for training. The highest precision and sensitivity models in both data sets were XG Boost-KNN. In contrast, the other ML methods are LR, SVM, and NB. LR predicts medical diabetes. Limitations include capturing complex variable linkages and interactions. LR's linearity assumption may make nonlinear correlations and interactions difficult. Skewed estimations may come from Un-independent medical datasets. Logistic regression can affect outliers. SVMs over fit due to noise and outliers. SVM performance depends on feature scaling, especially with different scales and distributions. Complex kernels and nonlinear decision constraints make large dataset SVM training computationally intensive. SVM performance depends on kernel and parameter selection, which is complicated and time-consuming. Interpretability is a difficulty in healthcare because prediction reasoning is crucial. NB, a probabilistic classifier, faces limitations in large, multidimensional datasets like medical diagnostics, including independence, insufficient data, and ignoring complex interactions like risk factors. The XG Boost-KNN evaluating matrices have high values.

5. Conclusion

Our investigation found that XGBoost and KNN beat other machine-learning methods by over 99 % in finding diabetic in its initial phases. The results of the study might possibly preserve

lives by helping medical professionals identify hyperglycemia early and make wise choices regarding therapy. Although we can adequately forecast diabetes, that our research has limits. The study's main drawback is the tiny sample size, making it challenging to validate any conclusions statistically. This model outperformed the other standard model with 94,87 % accuracy, 98,00 % precision, 99,04 % recall, and 96,85 % F1 score. We aim to gather more global data to improve disease categorization accuracy and precision. Next, in the dataset, we'll find further characteristics that might help identify diabetic complications on early.

References

1. Suryasa IW, Rodríguez-Gámez M, Koldoris T. Health and treatment of diabetes mellitus. International Journal of Health Sciences. 2021; 5(1). https://dx.10.53730/ijhs.v5n1.2864
2. Khan FA, Zeb K, Al-Rakhami M, Derhab A, Bukhari SA. Detection and prediction of diabetes using data mining: a comprehensive review. IEEE Access. 2021 Feb 12; 9:43711-35. https://dx.10.1109/ACCESS.2021.3059343
3. Tarumi S, Takeuchi W, Chalkidis G, Rodriguez-Loya S, Kuwata J, Flynn M, et al. Leveraging artificial intelligence to improve chronic disease care: methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. Forms of Information in Medicine. 2021 May 11; 60:e32-43. https://dx.10.1055/s-0041-1728757
4. Nomura A, Noguchi M, Kometani M, Furukawa K, Yoneda T. Artificial intelligence in diabetes management and prediction. Current Diabetes Reports. 2021 Dec; 21(12):61. https://dx.10.1007/s11892-021-01423-2
5. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019 Jan 1;165:292-9. https://dx.10.1016/j.procs.2020.01.047
6. Soni M, Varma S. Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (Ijert) Volume. 2020 Sep;9.
7. Yuvaraj N, SriPreethaa KR. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing. 2019 Jan 16;22(Suppl 1):1-9.https://dx.10.1007/s10586-017-1532-x
8. Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M. A fusion-based machine learning approach for the prediction of the onset of diabetes. InHealthcare 2021 Oct 18 (Vol. 9, No. 10, p. 1393). MDPI. https://dx.10.3390/healthcare9101393
9. Kaul S, Kumar Y. Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. SN Computer Science. 2020 Nov; 1(6):322.https://dx.10.1007/s42979-020-00337-2
10. Singh M, Bhambri P, Singh I, Jain A, Kaur EK. Data mining classifier for predicting diabetics. Annals of the Romanian Society for Cell Biology. 2021 Apr 17:6702-12.
11. Islam MM, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis 2020 (pp. 113-125). Springer, Singapore. https://dx.10.1007/978-981-13-8798-2_12
12. Rajendra P, Latifi S. Diabetes Prediction using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update. 2021 Jan 1; 1:100032. https://dx.10.1016/j.cmpbup.2021.100032
13. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering. 2021 Jun 1; 2:40-6. https://dx.10.1016/j.ijcce.2021.01.001
14. Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, et al. An AI-

based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. The Journal of Supercomputing. 2021 Feb; 77:1998-2017. https://dx.10.1007/s11227-020-03347-2

15. Al-Hameli BA, Alsewari AA, Alsarem M. Prediction of diabetes using hidden Naïve Bayes: a comparative study. InAdvances on Smart and Soft Computing: Proceedings of ICACIn 2020 2021 (pp. 223-233). Springer Singapore. https://dx.10.1007/978-981-15-6048-4_20