JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction Model Using Machine Learning

[1]Dhruv Garg, [2]Deepak Bhardwaj, [3]Dinesh Baghel, [4]Vanshika Gupta

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India

Abstract: In India alone, the number of diabetes patients has crossed the mark of 100 million. Many parameters are responsible for a person getting affected by diabetes like excessive cholesterol level, sugar level, irregular production of insulin in the body, and much more. People with diabetes mellitus can suffer from various life-threatening diseases such as kidney damage, loss of sight, regulation of heartbeat, etc. The research delves into the development and evaluation of a data-driven diabetes prediction system employing machine learning (ML) algorithms. To predict the risk of diabetes onset, the proposed method uses a variety of patient parameters, including demographics, lifestyle choices, and clinical markers. The project uses the Support Vector Machine (SVM) as the preferred model for predicting diabetes in a person. The preferred kernel for the SVM is Linear. The significance of responsible and effective model creation is emphasized by the thorough examination of ethical issues and privacy concerns of sensitive health data. Additionally, the paper addresses the interpretability of ML models to enhance comprehension for healthcare professionals and patients. The proposed diabetes prediction system exhibits promising accuracy and reliability, demonstrating its potential as a valuable tool for early diabetes risk assessment. These results provide a substantial contribution to the growing corpus of knowledge about machine learning uses for healthcare and provide insightful information about how to integrate these systems in real-world clinical settings. The authors have tested various machine learning models and found SVM as the most reliable reaching an accuracy of 76.62%. A website framework has been used to provide inputs and predict diabetes. The dataset is provided by the 'National Institute of Diabetes and Digestive and Kidney Diseases' of female patients [22].

Index Terms **- Support vector machine, diabetes mellitus, machine learning, early disease detection, Healthcare**

I. Introduction

Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin [1]. Insulin is a hormone produced by the pancreas that regulates blood sugar levels. Insulin's main function is to regulate the body's energy by balancing micronutrient levels. Diabetes Mellitus commonly known as diabetes has no exact cause but certain factors can lead a person to be more prone to being affected by diabetes like abnormal blood glucose levels, excess body weight, lack of physical activity, etc. Diabetes can lead to some of the very adverse effects on a person's life and can also lead to life-threatening situations if not properly cared for. These can include damage to the body's organs, and damage to large and small blood vessels, which can lead to heart attack, stroke, and problems with the kidneys, eyes, gums, feet, and nerves [2].

During the initial development of diabetes, it is difficult for medical professionals to diagnose and detect diabetes accurately in a person. Modern and advanced techniques such as Artificial intelligence (AI), deep learning, machine learning, and many more methods, can help medical experts have a better grasp of a person's condition in the early stages of diabetes and reduce the chance of being affected by diabetes and take a preventive measure from early stage to stop the impact of diabetes on the person's life. There have been many contributions by people worldwide to develop a technique that can predict diabetes automatically using machine learning. Most of the works use the Pima Indian dataset [3]. The dataset in the research was obtained from the UCI learning repository. The authors first pre-processed the dataset by cleaning, integrating, and reducing it. The accuracy level obtained was 90% using a random forest algorithm. In another paper [4], Aishwarya and Vaidehi used various prediction models such as Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, and many more. The authors compare the accuracies of different machine learning algorithms and find that the Logistic Regression and AdaBoost Classifier have the highest accuracy in predicting diabetes. Their results showed that the proposed model improves the accuracy and precision of diabetes prediction compared to existing methods. The authors suggest that this work can be extended to predict the likelihood of a non-diabetic individual developing diabetes in the future. The authors reached the result, having logistic regression achieving the accuracy of 96%. Debadri and Debpriyo [5] applied a diabetes Mellitus prediction system based on previous studies' results and their experiments with various prediction models. They concluded that the Random Forest algorithm outperforms Logistic Regression and Support Vector Machine algorithms in predicting diabetes, achieving an accuracy of around 84%. They also pointed out that Glucose level and age are identified as significant features for predicting diabetes. In conclusion, the researchers determined that Random Forest is the most effective algorithm for predicting diabetes. They recommend keeping glucose levels low, following a proper diet, and taking preventive measures for individuals with a family history of diabetes. The study highlights the importance of early detection and prevention of diabetes to avoid complications. The findings can contribute to improving diabetes diagnosis and management strategies. After reading multiple research papers we have concluded that the researchers have used a combination of multiple machine-learning models with different approaches. Most of the research focused on enhancing the accuracy level, which has prompted us to better evaluate our proposed system improve its reliability, training, and testing accuracy, and find a better model that caters to modern needs.

The trained machine-learning model has been deployed on the internet on a local server which is currently not available to the general public and is accessible as standalone as well as integrated into a website. The private dataset is females from Bangladesh, we used the Pima Indian Dataset in this paper [3]. The data had many missing values in some attributes, we replaced them with the mean value of each feature such as BMI and Age attributes as these attribute fields cannot have a value of zero. To train and test our machine learning model we have split the dataset using the holdout validation technique. The split is done in 80-20 parts, 80% of the data is kept for training the learning model while the remaining 20% dataset is kept for testing the accuracy of the trained model. In this research paper, we have applied a Support Vector Machine using a Linear kernel as it performed the best out of all the other learning models that we have tried in a different environment on the same dataset. Various algorithms were first tested in a rudimentary way multiple times on the same set of datasets, out of which all, the Support Vector Machine Linear kernel performed best when the average of all the algorithms and cases was calculated.

This research implements a diabetes mellitus prediction system using machine learning. The major contributions of the research are as follows:

- The main part of the work is to present a unique dataset of diabetes containing 768 samples. The private dataset has been provided by the 'National Institute of Diabetes and Kidney Diseases' and comprises female patients/personnel.
- Each record in the dataset has 7 attributes that are 'Pregnancies', 'glucose', 'blood pressure', 'skin thickness', 'insulin', 'BMI', 'Age', and 'Outcome' of diabetes.
- The missing values of attributes are filled by taking the mean of the respective attributes and replacing the missing values (zero in the field) with that mean.
- A standalone and website-integrated machine learning algorithm is deployed on the local server to make instantaneous predictions and provide results in real-time.

The Novelty of this research is to create and deploy a fully capable diabetes prediction system for a private dataset of female Bangladeshi patients using machine learning algorithms and provide results instantaneously.

The following section is a road map of the paper's structure The proposed diabetes prediction system has been discussed and illustrated in Section 2 with required figures and flowcharts. The final results of the research are presented in Section 3. Section 4 concludes the paper with some recommendations for future possible improvements.

II. Proposed System

This section briefly describes the working procedures and the implementation of the proposed machine learning algorithm i.e. SVM to design the discussed diabetes prediction system. Figure 1 below shows the different stages of the implementation of this system and the overall workflow of the system. The workflow diagram helps us better understand how the dataset is acquired, and the necessary steps it goes through to finally become a fully functional model fit for practical use. Let's discuss the workflow diagram briefly. Firstly, the dataset is acquired from the public domain such as the internet, the dataset in its current is called raw data or raw dataset is unfit for training the model. To make it usable for the machine learning model training it is pre-processed to remove any outliers, missing values, or any other values that may hamper or degrade the training results for example replacing null values with the mean values and eliminating outliers that greatly affect the relation between input and output. Then the dataset is separated into the training set and the testing set using the holdout validation technique in 80-20parts. 80% of the data is kept for training the model while the remaining 20% of data is used for testing the accuracy of the trained model. Afterward, the selected machine learning algorithm is applied to the training set to obtain a trained model and its accuracy is checked, if the obtained accuracy is found to be good enough for practical use the proposed website and standalone version of the training model are deployed, otherwise, the model has trained again with some performance tweaks and again check for the accuracy, this step is repeated until the desired accuracy is obtained.

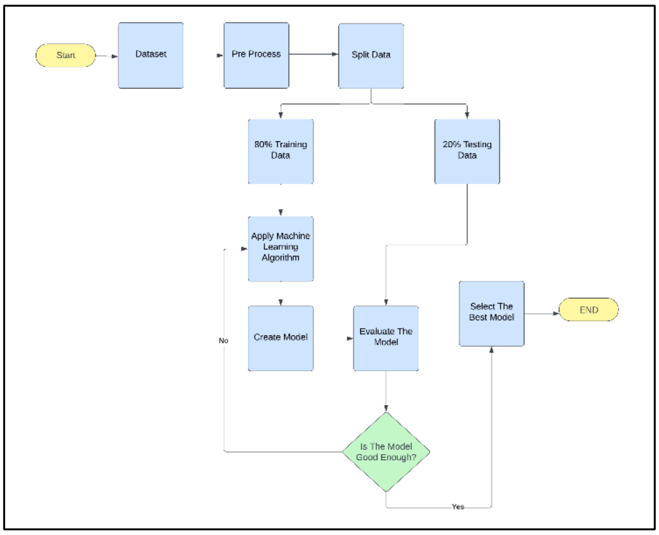


Fig.1: Flow diagram of proposed diabetes prediction system.

A. Dataset

The Pima Indian dataset is an open-source dataset [3] that is available to the general public for machine learning training and testing and is provided by the ‘National Institute of Diabetes and Kidney Diseases’. It contains 768 records of patient data out of which 268 patients have diabetes and 500 are not diabetic.

The dataset has the attribute ‘Outcome’ that indicates whether a patient is diabetic or not. If the value in ‘Outcome’ is 0 (zero) it means that the person is not diabetic and if the value in ‘Outcome’ is 1 (one), it means that the person is diabetic.

Figure 2 below shows the percentage of people having diabetes in the Pima India dataset.

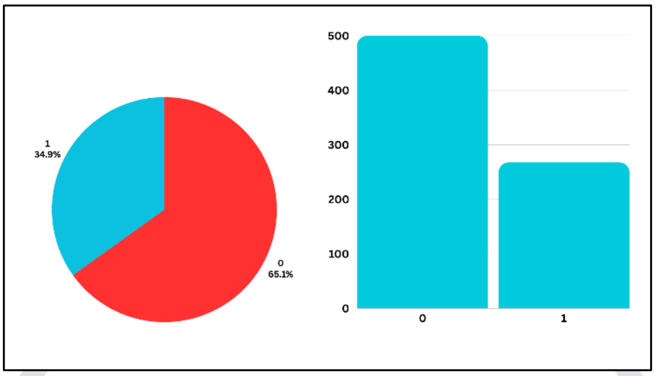


Fig.2: Percentage of people having diabetes in the dataset.

Table I below expresses the seven features of the open-source Pima Indian dataset not in a particular order.

Table I. FEATURES OF DATASET

Pregnancies	Skin Thickness	Glucose
Insulin	Age	BMI
Blood Pressure		

Table II below shows the minimum, maximum, and average of each of the attributes of the Pima India dataset. This helps us in having an overall estimate of the records of the patients in the dataset. To best know the data, it is advised to find the lowest, highest, and average values of each parameter.

Table II. ATTRIBUTES OF DATASET

Features	**Minimum**	**Maximum**	**Average**
Pregnancies	0	8	3.09
Skin Thickness (mm)	7	55.53	29.07
Glucose (mg/dL)	44	199	121.68
Insulin	14	410.61	152.30
Age (years)	21	72	33.21
BMI (kg/m^2)	21	68	33.21
Blood Pressure (mm Hg)	36.12	108.69	72.40

B. Dataset Preprocessing

In the acquired dataset, it is visible that some of the attribute values of multiple records have 0 (zero) to indicate null values but they cannot be zero, as it is practically impossible. For example, attributes like, ‘Skin Thickness’, ‘Age’, and ‘Body Mass Index’ (BMI) cannot be zero. The zero value of the same has been replaced by their respective attribute mean (average) values to ensure it does not adversely affect the ‘Outcome’ attribute.

The dataset is separated into two datasets, a training set and a testing set using holdout validation techniques in 80-20parts. 80% is used as a training dataset while the remaining 20% is used as a testing dataset. Equations 1 & 2 are used to calculate the upper limit of the attributes, to find out the outliers in the dataset.

Upper_limit=mean(attribute)_+_3*standard_deviation (attribute) (1)

Lower_limit=mean(attribute)_-_3*standard_deviation (attribute) (2)

Figure 3 below shows the correlation of various features, with each other through a heatmap. It describes the impact of each attribute in the dataset on all other attributes of the same. We can conclude from the below-given heatmap that ‘Age’ and ‘Pregnancies’ highly correlate with each other and ‘Skin Thickness’ and ‘BMI’ highly correlate with each other. There is a visual correlation between ‘Glucose’ and ‘Outcome’, indicating whether a patient is going to be diabetic or not. It is highly dependent on the amount of glucose in the blood level of the person.

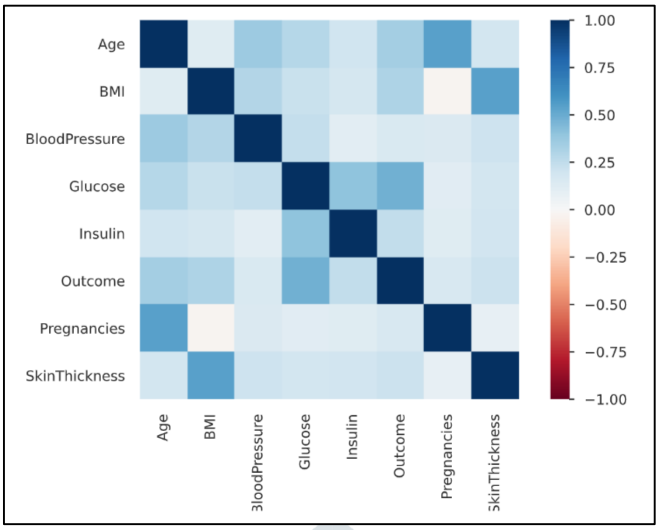


Fig. 3: Individual features impact each other.

C. Machine Learning Algorithm

In this research, multiple machine learning algorithms and techniques have been tested for the implementation of the diabetes prediction system, in a rudimentary way, and is collectively decided by the authors to only use and show the algorithm that performed best in the selection stage to be implemented in the final model.

Table III below shows the machine learning algorithm tested and their accuracy obtained while keeping the training and testing data in the same ratio which is 80:20 respectively. As seen in the table, SVM Linear performs the best and is selected for the final implementation of the prediction system.

Table III. ALGORITHMS ACCURACY

Algorithm	Accuracy
SVM – Sigmoid Kernel	74.67 %
SVM – RBF Kernel	66.23 %
SVM – Linear Kernel	**76.62 %**
Logistic Regression	74.67 %
Random Forest Classifier	73.37 %

Support vector machine (SVM): SVM performs supervised classification by choosing the best hyperplane [6]. In the following study, we tried various SVM kernels in the training set, finally, we found that the SVM with a linear kernel produces the best results in the Pima India dataset [7].

D. Deployment Of Prediction System

The proposed machine learning diabetes prediction system has been deployed as a standalone and as a website version on a local server, which is fully functional as intended. The system is capable of predicting diabetes in a person instantaneously with real data given by the user. However, the system is not made available to the general public and is kept for use only by authorized personnel on the local server.

Standalone version: We have used Streamlit and Python to deploy and develop the predicting system respectively. Streamlit provides us with an easy frontend need which is the UI through which the user can interact and enter the data. This uses Anaconda to provide the Python environment to run in offline mode.

Web application: We have used HTML, and CSS to develop the frontend website on which the prediction model is hosted. We also used the Streamlit deployment feature to directly integrate the standalone version on one of the pages of our websites without the need for any hard coding. The Microsoft Vs Code was used as the code editor.

III. Results and discussions

This section discusses the results and overview of the proposed diabetes prediction system. In Figure 4, we can see how each attribute of the open-source Pima India dataset affects the 'Outcome' attribute, which determines whether a patient is diabetic or not. From seeing the bar graph, we can see that 'Glucose' attributes greatly correlate and affect the value of the 'Outcome' Attribute, followed by 'BMI' and 'Age'.

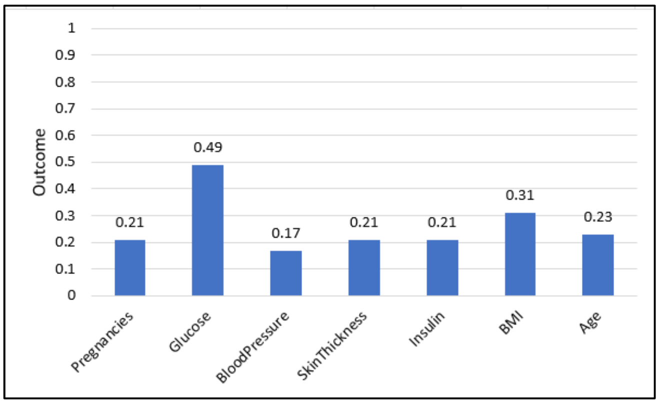


Fig. 4: Features hierarchy concerning diabetes.

Figure 5, represents the SVM-Linear graph of the trained machine learning model, and the blue line across the graph represents the hyperplane of the graph, classifying the data into two classes, that are diabetic or not diabetic.

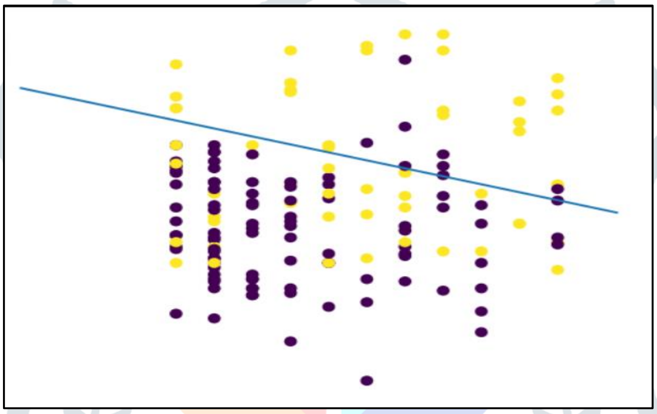


Fig.5: Classified data points.

Figure 6 represents the Receiver Operating Characteristics (ROC) curve, the x-axis represents the False Positive Rate (FPR) and the y-axis represents the True Positive Rate (TPR). The value obtained is 0.81, which is considered good performance, suggesting a better discrimination and classification performance.

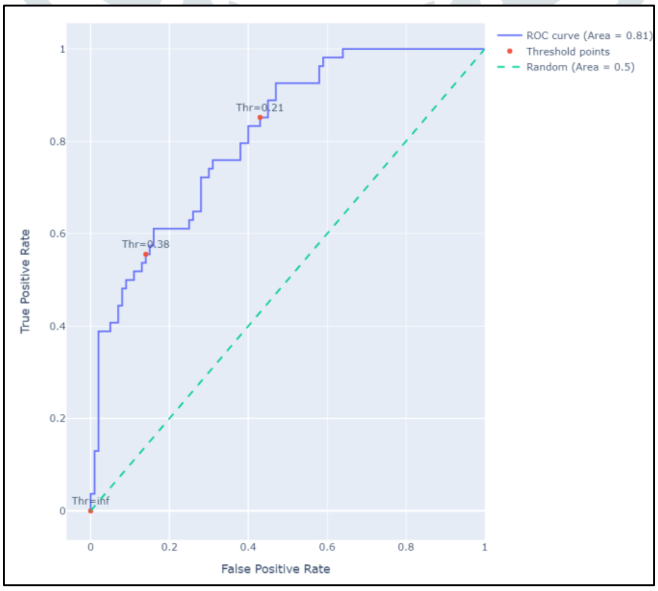


Fig. 6: ROC curve.

Figure 7 represents the confusion matrix of the trained linear support vector machine. Using the confusion matrix we can conclude the precision score and f1 score.

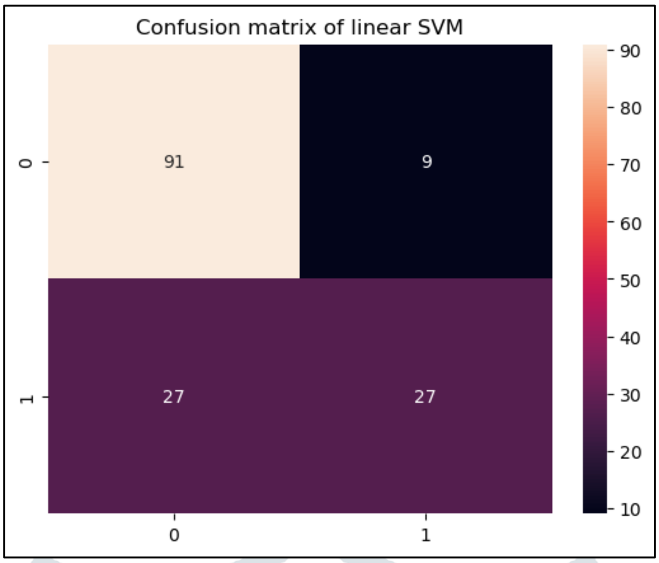
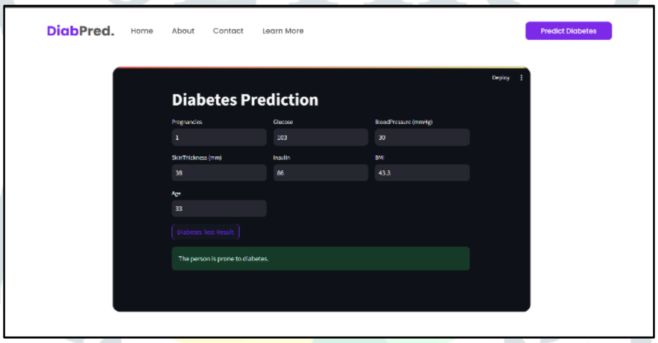


Fig.7: Confusion matrix.

Figure 8 shows the page of the website in which the trained machine learning model is integrated and deployed. This is the page that will be visible to the user and will act as a GUI for the user to insert their data.



Diabetes Prediction

Pregnancies: 1
Glucose: 103
BloodPressure (mmHg): 30
SkinThickness (mm): 38
Insulin: 86
BMI: 43.3
Age: 33

The person is prone to diabetes.

Fig.8: Learning model implementation.

IV. Conclusions

Diabetes can become a factor that may cause a reduction in life expectancy and complications of many diseases in the long run. In this research, we have proposed and developed a diabetes prediction system using machine learning, which produces instantaneous results from real data and achieves an accuracy score of 76.62%. The open-source Pima India dataset [7] has been used in this paper. The dataset was pre-processed to remove any null values and any outliers to train the model more precisely. The system trains the model with a Support Vector Machine with a Linear Kernel. The proposed system is implemented in two different forms, one is a standalone version and the other is a website in which the machine learning model is integrated. The future scope of this model, for example, can be acquiring a private dataset with a larger sample size and more attributes having precise values to obtain better results. Another addition could be using more complex algorithms that specifically cater to large and multi-attribute datasets.

V. Acknowledgment

We would like to express our gratitude to everyone who contributed to the successful completion of this project and research paper. We would like to thank our project supervisor, Ms. Vansika Gupta, for their invaluable guidance, and constructive feedback throughout the entire research process. We are also thankful to the faculty members of the Department of Computer Science and Engineering (IPEC) for their encouragement and for providing a conducive academic environment that allowed me to explore and develop my research interests. We are grateful to IPEC affiliated with DR APJ Abdul Kalam Technical University for providing us with this opportunity to work.

References

[1] A. T. a. H. M. D. Kharroubi, "Diabetes mellitus: The epidemic of the century.," World journal of diabetes, vol. 6.6, p. 850, 2015.

[2] "Better Health," 23 03 2021. [Online]. Available: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.

[3] J. W. J. E. E. a. W. C. D. Smith, "bowler, WC, and Johannes, RS 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Symp. Computer Appl. Med. Care 1988 Proc, Chicago, 1988.

[4] A. a. V. V. Mujumdar, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, 2019.

[5] D. D. P. a. P. G. Dutta, "Analysing feature importances for diabetes prediction using machine learning," Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018.

[6] J. G.-L. F. R.-M. L. a. L. A. Cervantes, "A comprehensive survey on support vector machine classification: Applications, challenges and trends.," in Neurocomputing, 2020, pp. 189-215.

[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[9] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[10] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[11] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm", Journal Of Computing, Volume 3, Issue 12, December 2011.

[12] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[13] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[14] Stone MA, Camosso-Stefinovic J, Wilkinson J, de Lusignan S, Hattersley AT, Khunti K. Incorrect and incomplete coding and classification of diabetes: a systematic review. Diabet Med. 2010; 27:491–497.

[15] Vehik K, Hamman RF, Lezotte D, Norris JM, Klingensmith GJ, Dabelea D. Childhood growth and age at diagnosis with Type 1 diabetes in Colorado young people. Diabet Med. 2009; 26:961–967.

[16] Elmarakby AA, Sullivan JC. Relationship between oxidative stress and inflammatory cytokines in diabetic nephropathy. Cardiovasc Ther. 2012; 30:49–59.

[17] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020.

[18] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019.

[19] Herron P., "Machine Learning for Medical Decision Support: Evaluating Diagnostic

[20] P. Saeedi et al. [Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas] Diabetes Res. Clin. Pract. (2019). Performance of Machine Learning Classification Algorithms", INLS 110, Data Mining, 2004.

[21] Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2022). Identification of bacterial cell wall lyases via pseudo amino acid composition. Biomed. Res. Int. 2016:1654623. doi: 10.1155/2016/1654623

[22] U.M.L.(Owner),"Kaggle,"[Online].Available:https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data.