ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

DIABETES PREDICTION USING SUPPORT VECTOR MACHINES

N. Srividhya[1], K. Divya[1], N. Sanjana[1,] K. Krishna Kumari[1], M. Rambhupal[1]
[1]Department of Information Technology, Vasireddy Venkatadri Institute of Technology Guntur.

Article DOI: https://doi.org/10.36713/epra14769
DOI No: 10.36713/epra14769

ABSTRACT
One of the worst illnesses in the world is diabetes. It is also the creator of a variety of other diseases, such as urinary organ illness, blindness, and cardiac failure. The patient must go to a diagnostic facility in this situation to receive their reports following consultation. Because of this, they always have to invest both money and time. However, as machine learning techniques have improved, we now have the freedom to look for the right solution. For example, we now have sophisticated information processing systems that can predict whether a patient has polygenic disease or not. Additionally, anticipating the illness early results in giving the patients what they need before it becomes urgent.Goals of this analysis is to develop a system which predicts the diabetes risk level of patient.The experimental results shows that the prediction of diabetes done at high accuracy using support vector machines.

KEYWORDS: ***Early Detection, Machine Learning, SVM(Support Vector Machines), Accuracy.***

INTRODUCTION

Diabetes is a common chronic disease. Diabetes can be identified when blood glucose is higher than normal level, which is caused by high secretion of insulin and biological effects. Diabetes can cause various damage to our body and can disfunction tissues, kidneys, eyes and blood vessels The identification of such chronic disease at the beginning period could help specialists around the globe in forestalling loss of human life.

Diabetes can be divided into two categories, type-1 and type-2 diabetes Patients with type-1 diabetes are normally younger with n age less than 30 years old.
The clinical symptoms increase thirst and frequent urination.

This type of diabetes cannot be cleared by medications as it requires therapy. Type-2 diabetes occurs more commonly on middle-aged and old people, which can show hypertension, obesity and other diseases.

Being one of the main causes of mortality is diabetes mellitus. The current need is for early diabetes detection and diagnosis. An major categorization issue is the diagnosis of the diabetes disease and the interpretation of the diabetes data

It is necessary to create a classifier that is accurate, practical, and cost-effective. A lot of human ideologies are provided by artificial intelligence and soft computing techniques, which are also used in human-related domains of application. These systems are useful for making diagnoses in medicine.

The main topic of this research report is "Diabetes Detection Using Support Vector Machines (SVM)," a sophisticated machine learning technique that has become well-known for its efficiency in resolving challenging classification issues. SVM's capacity to manage high-dimensional data and spot subtle trends within datasets makes it especially well-suited for medical diagnosis applications, such as diabetes detection. However, with the remarkable growth of Machine Learning and advanced information processing techniques, we now have the tools to address this health crisis effectively. These cutting-edge techniques grant us the ability to predict the onset of polygenic illnesses such as diabetes with a level of precision and efficiency previously deemed unattainable. Additionally, the proactive forecasting of illnesses is the first crucial step toward providing timely intervention, potentiating the potential for a cure.

The accuracy of various different approaches employed for the diabetes categorization dataset ranged from 59% to 77.5%. SVMs have demonstrated impressive performance when using Computer Aided Diagnostic (CAD) systems to enhance diagnostic choices . Vapnik was the first to introduce the Support Vector Machine (SVM), a unique learning device that has lately been used in a number of financial applications, primarily in the field of time series prediction and classification.

SUPPORT VECTOR MACHINE

SVM Model Generation
SVMs are also employed because they can identify intricate associations in your data without requiring you to perform a lot of manual modifications. When working with smaller datasets that contain tens to hundreds of thousands of characteristics, it's an excellent choice. Because they can handle small, complex information better than other algorithms, they usually find more

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

accurate findings.

SVM is a group of related supervised learning techniques used in regression and classification diagnostics in medicine [1,16]. SVM maximizes the geometric margin while also minimizing the empirical classification error. Therefore, SVM stands for Maximum Margin Classifiers. The Structural Risk Minimization Principle, or assured risk boundaries in statistical learning theory, forms the foundation of SVM, a general algorithm. Using an implicit mapping of their inputs into high-dimensional feature spaces, known as the kernel technique, support vector machines (SVMs) may effectively execute non-linear classification. The classifier may be constructed without explicitly knowing the feature space thanks to the kernel trick.An SVM model is a mapping of the instances as points in space, so that the examples belong to distinct categories and are separated by as large a distance as feasible [1, 8]. An SVM, for instance, locates a hyperplane with the highest proportion of points from the same class on the same plane given a set of points that belong to either of the two classes. The optimal separating hyperplane (OSH) is a separating hyperplane that minimizes the possibility of misclassifying test dataset samples while optimizing the distance between the two parallel hyper planes.

Given labeled training data as data points of the form
$M=\{(x_1,y_1),(x_2,y_2), \ldots . ,(x_n,y_n)\}$
Where y_n=1/-1, a constant that denotes the class to which that point x_n belongd.The variable n denotes the number of samples.Each x_n denotes a p-dimensional real vector.

The SVM classifier performs classification using an appropriate threshold value after first converting the input vectors into a decision value. The hyperplane is divided (or separated) in order to display the training data. This can be explained as follows:
Mapping : $w^T.x+b=0$
Where w is the weight of vector i.e., perpendicular to the hyperplane.
x is the feature vector of a data point
b is the bias term
Minimize $½||w||^2$ ensure for all i,
$y_i(w^Tx_i+b)>=1$
Where :
y_i is the class label of the datapoint.

The data points that are closest to the hyperplane are called support vectors, and it is these that establish the margin. When the support vectors are evenly spaced and properly categorized, the margin is maximized relative to the hyperplane.

Because SVMs seek to identify a hyperplane that optimizes the separation between classes, they are an extremely strong tool for binary classification. This approach also makes SVMs resistant to outliers and very effective in a wide range of real-world scenarios.

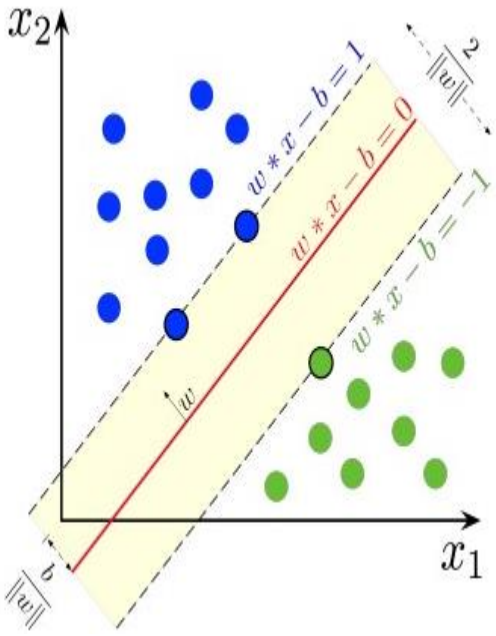


Fig.1.Maximum margin hyperplanes for SVM trained with samples from two classes

Existing System

Algorithms including k-Nearest Neighbors (KNN), SVM, Decision trees, LR, Random forests, and Naïve Bayes are used to predict diabetes.The comparative analysis is performed using a standard diabetic dataset.A number of pregnancies, blood pressure, glucose, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome are among the attributes that are used in the current dataset.Under the current system, only female patients who are at least 21 years old are diagnosed.Existing systems' efficacy varies, so before implementing them, it's critical to verify their accuracy using pertinent datasets, which adds to the implementation's time complexity.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

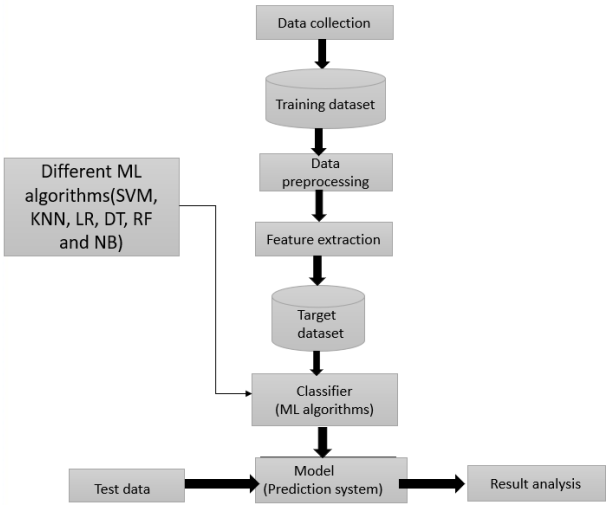


Fig.2.Existing System Architecture

Proposed System

We use a single algorithm in the proposed system, which lowers the time complexity.SVM (Support Vector Machine) is a machine learning technique used to predict diabetes.We are able to take into account patient data regardless of age or gender.The suggested system is an interactive application that asks the user to enter data in order to generate a prediction.

The updated dataset under consideration includes the following attributes: gender, age, heart disease, hypertension, smoking history, BMI, hemoglobin A1c (HbA1c) level, glucose level, and outcome.The proposed system takes into account patients who are younger than 21.

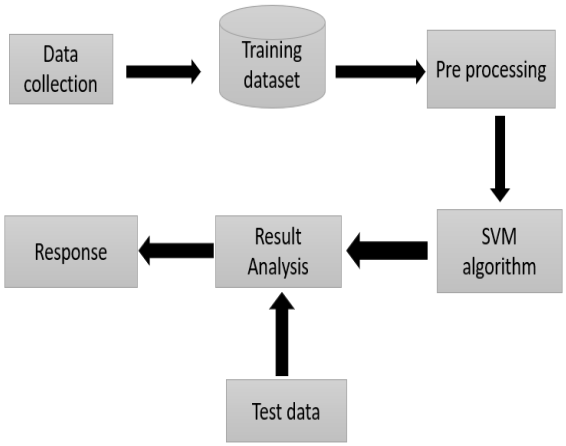


Fig.3.Proposed System Architecture

DISEASE CLASSIFICATION USING SVM

A. Experimental Setup

The diabetes dataset has led to the development of SVM models for classification. The Matlab R2010a is used to conduct the studies. The datasets are read straight out of Matlab and saved as MS Excel documents. The Receiver Operating Characteristic (ROC) curve is used to assess the created models' diagnostic performance. Plotting the genuine positive rate (sensitivity) against the false positive rate for various cut-off values is how the ROC curve is made. A sensitivity/specificity pair that corresponds to a specific decision threshold is represented by each point on the ROC plot.

B. Diabetes Disease Dataset

The dataset for the suggested system includes both numerical and characteristic data. All patients can have the prediction made, regardless of their age or gender.The dataset, which has more than two lakh patient records, may be utilized for testing as well as training."0" or "1" are the possible values for the binary target variable. "0" denotes a negative test result for diabetes, while "1" indicates a positive result. Maximum of cases are in class "0" while some of thecases are in class "1". By fine-tuning parameters, the relevance of the automatically selected collection of variables was assessed further by hand. The variables that performed the best in terms of discrimination were those that made the final cut.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

Eight variables—numeric and characteristic—are present: (1)Gender, (2)an oral glucose tolerance test's plasma glucose concentration after two hours(HbAlc_level), and (3) Blood pressure diastolic (mm Hg),(4) Heart disease, (5)smoking history, (6)body mass index(BMI), (7)hypertension and (8)age (years).Even when there are missing values in the data set, these can be handled by other processes.One of the characteristic variables is smoking history, which accepts inputs in the following three formats: never, no information, and current.

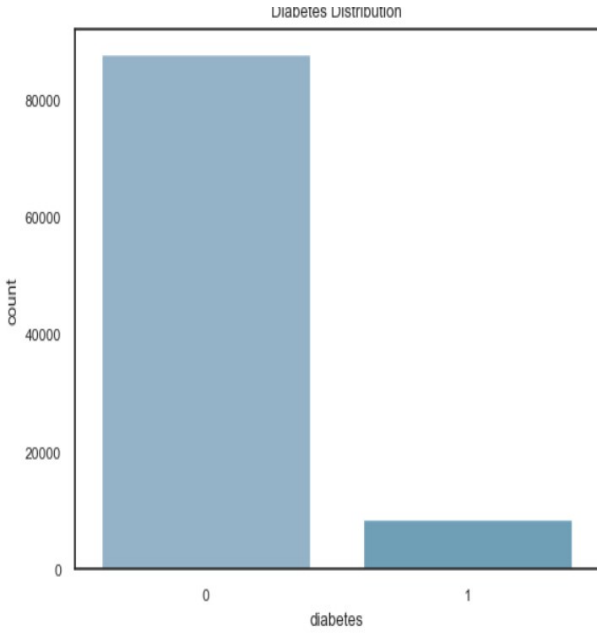
	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.00	0	1	never	25.19	6.60	140	0
1	Female	54.00	0	0	No Info	27.32	6.60	80	0
2	Male	28.00	0	0	never	27.32	5.70	158	0
3	Female	36.00	0	0	current	23.45	5.00	155	0
4	Male	76.00	1	1	current	20.14	4.80	155	0

C. Training and Test dataset evaluation

In the training data set, a 10-fold cross-validation was carried out to assess the SVM models' resilience. Ten equal-sized subgroups are initially created from the training data set. A model trained on all cases and an equal number of non-cases randomly chosen from the remaining nine datasets was trained using each subset as a test data set. Ten iterations of this cross-validation procedure were conducted, with one test data set serving as each subset. Test data sets evaluate the models' performance.

RESULT ANALYSIS

When assessing the effectiveness of a classification model—like the one used to predict diabetes—a confusion matrix is a useful tool. The confusion matrix in a binary classification problem such as this one (diabetic or non-diabetic) usually has four elements:



True Positives (TP): The quantity of diabetes patients who were accurately predicted.
True Negatives(TN):The quantity of appropriately predicted non-diabetic patients
False Positives (FP): The quantity of individuals who are not diabetic but are mistakenly diagnosed as such.
False Negatives (FN): The quantity of patients with diabetes who are mistakenly labeled as non-diabetic.

Table I

Data Set	Samples	Training Data	Testing Data	Attributes	No. Of Classes	Accuracy
Diabetes	1 lakh	80,000	20,000	8	2	96.5

Accuracy = (TP + TN) / (TP + TN + FP + FN)

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

```
# Create an instance of the SVM classifier
svm_model = SVC(kernel='rbf', random_state=42)

# Train the SVM model
svm_model.fit(X_train_scaled, y_train)

# Make predictions on the test data
y_pred = svm_model.predict(X_test_scaled)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:",  round(accuracy_score(y_test, y_pred) * 100, 1))
```

```
Accuracy: 96.0
```

To comprehend your SVM model's decision boundary, create visualizations. This can assist you in understanding how the model divides data points into various classes.

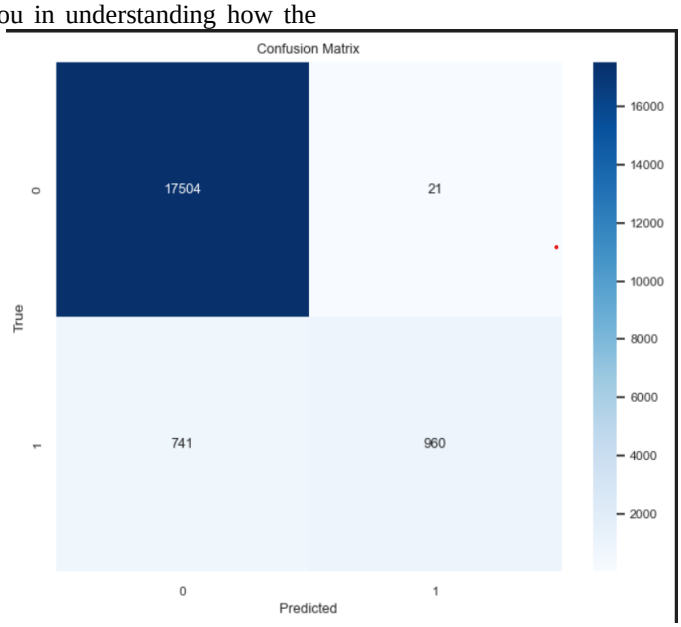


Fig Confusion Matrix

This project's experimental results show a remarkable degree of efficiency and accuracy, which is very promising. The ability of the SVM-based model to reliably predict diabetic risk levels has been demonstrated, and this can greatly enhance patient outcomes. Healthcare professionals can rely on the system as it reduces the possibility of false positives and false negatives with high accuracy.

CONCLUSION

In a time when technology and healthcare are increasingly combining, the creation of a diabetic risk level prediction system has shown promise for improving patient care and overall health. Diabetes is a worldwide health issue that has far-reaching effects. It frequently results in serious complications such as kidney disease, blindness, and heart failure. Patients have a significant time and financial burden because frequent trips to diagnostic centres become essential. The creation of this SVM-based diabetes risk prediction system is a major step in the direction of more efficient and proactive healthcare. One way to lessen the overall healthcare burden associated with complications related to diabetes is to be able to provide patients with timely information and interventions. The project's success highlights machine learning's potential in the healthcare industry and the beneficial effects it can have on people's lives all over the world. The feature subset selection procedure can be used in the future to enhance the SVM classifier's performance.

REFERENCES

1. *Rambhupal, M., Voola, P. An effective hybrid attention capsule autoencoder model for diagnosing COVID-19 disease using chest CT scan images in an edge computing environment. Soft Compute (2023). https://doi.org/10.1007/s00500-023-09111-x*
2. *Ramesh, J., Aburukba, R., Sagahyroon, A.: A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technol. Lett. 8, 45–57 (2021) [PMC free article] [PubMed] [Google Scholar]*
3. *Yu, L., Liu, L., Peace, K.E.: Regression multiple imputation for missing data analysis. Stat. Methods Med. Res. 29, 2647–2464 (2020)*

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 9| Issue: 10| October 2023|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

4. *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.*
5. *B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7, 2017.*
6. *P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.*
7. *Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y.: AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J. Supercomput. 77, 5198–5219 (2021)*
8. *Mounika, V., Neeli, D.S., Sree, G.S., Mourya, P., Babu, M.A.: Prediction of type-2 diabetes using machine learning algorithms. In: International Conference on Artificial Intelligence and Smart Systems, pp. 127–131 (2021)*
9. *Tran, C.T., Zhang, M., Andreae, P., Xue, B., Bui, L.T.: Multiple imputation and ensemble learning for classification with incomplete data. In: Intelligent and Evolutionary Systems; New York: Springer, pp. 401–415 (2017)*
10. *P. Cıhan and H. Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in 2021 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.*
11. *F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater. Today, Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.*
12. *H. EL Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "Ontology-Based Machine Learning to Predict Diabetes Patients," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978-3-030-91738-8_40.*
13. *Z. Sabouri, Y. Maleh, and N. Gherabi, "Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 173–179. doi: 10.1007/978-3-030-91738-8_17.*