



Predictive Modelling for Diabetes: A Support Vector Machine Approach

Hemanth R¹, V Tushar², Suchit Karnam³, Shravani G Kamath⁴, Pallavi K V⁵

^{1,2,3,4} Department of Computer Science and Engineering, AMC Engineering College,
Bengaluru, India

⁵ Assistant Professor, Dept. of Computer Science and Engineering, AMC Engineering College,
Bengaluru, India

ABSTRACT

A chronic disorder that affects blood glucose levels, diabetes mellitus is caused by either cellular resistance or inadequate insulin and can lead to problems such as renal and heart disease. Support Vector Machines (SVM) are investigated in this work for the prediction of diabetes, with an emphasis on clinical application, model validation, and performance measures. SVM exhibits good prediction capabilities when used with feature selection methods like PCA and kernel functions like RBF. The performance of hybrid models, such GA-SVM, is further improved. SVM's ability to classify high-dimensional data offers promise for early diabetes detection; nonetheless, regularization and data imbalance management need to be improved.

Keywords: Machine Learning, SVM, Health Care, Genetic Algorithms, Kernel Functions (Linear, Polynomial, RBF), Feature Selection (PCA, RFE), K-fold Cross-Validation

1 Introduction

Diabetes Mellitus is a chronic medical condition because the body's mechanism to maintain appropriate sugar in the blood, has failed. The hormone insulin is one produced by the pancreas, which helps cells inside the body absorb glucose. Either it is not produced sufficiently or the cells develop a resistance to insulin function. Severe health consequences of continued high levels of blood glucose include Heart diseases, Kidney diseases, Nerve problems and eye disorders. There are 2 kinds of Diabetes Mellitus. Type 1 Diabetes is an autoimmune disease wherein the immune system attacks the pancreatic cells, responsible for the production of insulin, inappropriately. Patients with Type 1 Diabetes must be given daily doses of insulin. With Type 2 Diabetes, the body either stops making enough amounts of insulin or becomes insensitive to its presence. This is the most common one and it is treated with insulin, oral drugs and changes in lifestyle. Gestational Diabetes often occurs during pregnancy but tends to disappear after delivery unless the mother develops Type 2 Diabetes later in her life.

Heart diseases and stroke are strongly related to Diabetes Mellitus as well as drastically raise heart disease risks. Hyperglycemia damages nerves in the limbs that cause diabetic neuropathy, which extreme leads to amputation. Diabetic neuropathy or damage to the kidney is also the most common cause of renal failure and chronic disease of the kidney. In diabetic retinopathy, the blood vessels of the retina get damaged that lead to the loss of visions or complete blindness.



SVM is one of the good supervised machine learning algorithm that are particularly useful for classification problems especially for high dimensional domains. This algorithm finds a hyperplane in the given data, which best classifies the different classes. SVM can assess several medical factors to classify whether an individual is diabetic or non-diabetic. Even studies reveal that SVM surpasses many other algorithm in this regard, thus promising that this algorithm will make predictions as accurate as possible.

2 Prediction Factors

Glucose: - Insulin resistance brought on by glucose affects the body in type 2 diabetes, resulting in elevated blood sugar levels. Fatigue is the result of cells not absorbing enough energy. Chronically high blood sugar levels can cause major side effects over time, such as renal failure, nerve damage, and cardiovascular disease. Controlling blood sugar levels is crucial.

Insulin: - Insulin resistance reduces the efficiency of insulin in type 2 diabetes mellitus, which stops glucose from entering cells to be used as fuel. The pancreas initially makes up for this by generating more insulin, but eventually it might not be able to keep up with demand. High blood sugar levels and related problems are caused by this imbalance.

BMI: - Body mass index (BMI) has a significant influence on type 2 diabetes mellitus because high BMI is often linked to increased insulin resistance. Because extra body fat, particularly around the abdomen, disrupts insulin function, blood sugar levels may rise. Maintaining a healthy BMI via diet and exercise can help prevent or control diabetes.

Blood Pressure: - Because raised body mass index (BMI) is often linked to increased insulin resistance, BMI has a significant influence on type 2 diabetes mellitus. The disruption of insulin function caused by extra body fat, particularly around the abdomen, might result in elevated blood sugar levels. By eating well and exercising, one can maintain a healthy BMI and avoid or control diabetes.

Skin Thickness:- Insulin absorption rates in diabetics can be influenced by skin thickness, especially in regions such as the thighs and belly. Blood sugar regulation may be impacted by delayed absorption caused by thicker skin. Skin thickness and diabetes care, however, have a complicated and individual-specific interaction. A healthcare expert should always be consulted for individualized guidance on managing diabetes and administering insulin.

Diabetes Pedigree function: - An indicator of a person's hereditary susceptibility to diabetes is their diabetes pedigree function (DPF). It considers the age of onset and if close relatives have diabetes. An increased risk of diabetes is indicated by a higher DPF, particularly when paired with other risk factors including as obesity and a sedentary lifestyle. It's crucial to remember, though, that DPF is only one of several factors that increase the risk of developing diabetes.

Pregnancy: - Gestational diabetes may develop during pregnancy, or pre-existing diabetes may worsen. The body has a tougher time controlling blood sugar during pregnancy due to hormonal changes. Birth defects, preterm birth, and high birth weight are among the issues that may result for both mother and child as a result of this. To reduce these risks, careful blood sugar monitoring and control are essential throughout pregnancy.



Age: - The chance of getting type 2 diabetes is greatly increased with age. Our bodies lose their sensitivity to insulin as we age, which makes controlling blood sugar levels more difficult. Moreover, obesity, high blood pressure, and high cholesterol are among the other health issues that are more common in older persons and can exacerbate diabetes. Although it can strike at any age, children and young people are the ones who are diagnosed with type 1 diabetes the most frequently.

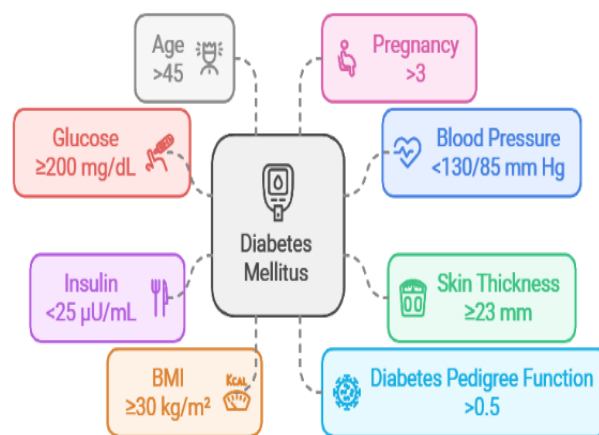


Figure 1 Factors Effecting Diabetes Mellitus

3 Objectives

To Evaluate the Effectiveness of SVM in Diabetes Prediction

It is going to be an experiment to find out how good a Support Vector Machines (SVM) is used for diabetes prediction, and it is about running some of the performance measures such as accuracy, sensitivity and specificity. This work is going to be applying SVM to train and test diabetes datasets to see how accurate it can be in distinguishing between diabetic cases and non-diabetic cases. By comparing SVM's predictive indices such as accuracy and F1-score to those of other models, researchers can evaluate the real-world clinical and research utility of SVM in the field of diabetes prediction.

To Validate the Model Using Real-World Data

This task is to use this SVM model to real world diabetes clinical or population based. Validation will be a model's performance on novel, previously unobserved data, and will be used to assess its accuracy, robustness and generalisability away from the training environment. By testing rates of accuracy, sensitivity and specificity in the field, this objective ensures that the SVM model is robust and generalises well to the prediction of diabetes in the field for a variety of patient populations and clinical contexts.[1]

To Provide Recommendations for Clinical Application

This goal is to give actionable suggestions for the deployment of the SVM model in the clinical



setting for the prediction of diabetes. Clinical recommendations will be made regarding the use of SVM in early diagnosis of diabetes based on the performance, practicality, and interpretability of the models. System requirements, model retraining, data confidentiality and usability will all be taken into account. These proposals aim at bridging the gap between research and practice, and at offering a framework that could be used to extend the SVM model to the real world, in a clinical setting.

4 Support Vector Machine

4.1. Effectiveness of SVM in Medical Prediction Models:-

It is now quite popular in medical prediction tasks since it can classify complex high-dimensional data (for example, patient health data). SVM has been a popular application in diabetes prediction because of the strength of SVM and its non-linear classification capability using kernel functions. [2]

4.2. Kernel Functions for Improved Performance:-

- The kernels commonly employed in SVM for diabetes classification are linear, polynomial and Radial Basis Function (RBF). Experiments indicate that RBF kernels are quite effective for medical data because of the flexibility they provide to model non-linear relations.
- To predict diabetes accurately, the right kernel is necessary. Empirical studies have shown that RBF and polynomial kernels work better than linear kernels in most real-world applications of the non-linearity of diabetes datasets. [3]

4.3. Feature Selection and Dimensionality Reduction:-

- In diabetes prediction models, the use of feature selection techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) is incorporated to identify the relevant features in the dataset and reduce the dimensionality of the data being used without sacrificing too much predictive power
- SVM (and PCA with SVM) are so much better because they remove the noise and leave only the features with predictive weight and also the speed and accuracy of the model. [4]

4.4. Performance Metrics and Validation Techniques:-

- Common metrics used for evaluating SVM in diabetes prediction are accuracy, sensitivity and specificity, F1 score, and area under the curve (AUC).
- Because in all SVM models, K-fold cross-validation is always applied to avoid overfitting so that the model will generalize to other unseen diabetes datasets. [5]

4.5. Hybrid Models Incorporating SVM:-

- SVM is usually applied together with other algorithms, such as Genetic Algorithms (GA) or Decision Trees, to enhance the prediction accuracy and robustness of the model in the area of diabetes prediction.



- It is suggested in several reports that GA-SVM hybrid models are more accurate due to their capability to optimize the selection of features and SVM parameters, and thus are highly appropriate for the diagnosis of diabetes.[6]

4.6. Comparison with Other Machine Learning Techniques:-

- SVM models have been compared with neural networks, Naive Bayes, and decision trees for diabetes prediction. While SVM models have been contrasted with NN, NB and DT in diabetes prediction. Although neural networks might beat SVM on some of the large datasets, SVM is usually more understandable and less computationally intensive to train.[7]

4.7. Real-World Applications and Case Studies:-

- Several case studies in the literature have demonstrated the effectiveness of SVM-based models in predicting the onset of Type-2 diabetes from the patient's history of illness, diet and exercise patterns.
- SVM has found its way into clinical decision support systems for early detection of diabetes, helping clinicians identify at-risk patients.[8]

4.8 Challenges and Future Directions:-

- Class imbalance in diabetes datasets, the extensive parameter tuning, and computational complexity are some of the major drawbacks of SVM.
- Current research work aims to minimise computational expense, increase interpretability, and augment SVM's ability to deal with large, imbalanced datasets using methods such as resampling and synthetic data synthesis.[9]

5 Workflow

The diabetes prediction pipeline in deep learning is a series of data-driven steps to create a model that, given certain variables, can predict the likelihood of diabetes. Here's an outline of each step: **Data Collection:** This first step of laying the foundation will be to collect a dataset pertaining to diabetes prediction. That could encompass medical history, lifestyle, genetics and other biomarkers. For this kind of work, datasets like Pima Indians Diabetes Dataset are commonly employed (Kaur et al, 2019).

Feature Selection: After data collection, the next step is identifying the most relevant features. Proper feature selection can reduce the model's complexity and improve its performance by using only those attributes that are most relevant to predicting the risk of diabetes, including age, blood pressure, body mass index (BMI) and glucose levels (Sharma Saini, 2020).

Feature Extraction: After identifying important features, feature extraction maps these attributes into a form amenable to modelling. This step might entail scaling, normalising or otherwise transforming features to make them suitable for deep learning algorithms (Sarwar et al, 2018).

Training: The training process consists of pumping the model up with curated and modified data. In this stage, the deep-learning algorithm is trained to recognise the patterns in the data and to distinguish diabetic from non-diabetic subjects. Training frequently uses techniques such as



Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) (Sahu et al., 2021).

Testing: Testing evaluates the model's accuracy using test data. Accuracy, precision and recall are performance metrics that are used to estimate the model's ability to generalize on unseen data and thus provide trustworthy diabetes prediction (Kumar et al., 2021).

Diabetes Prediction Process Using Deep Learning

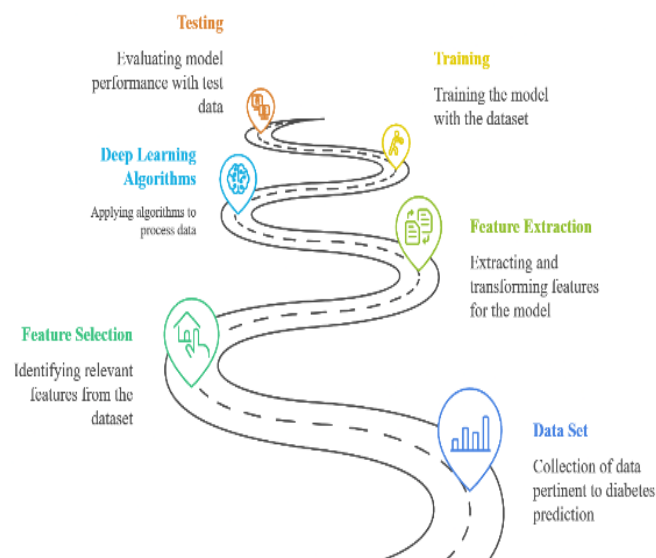


Figure 3: Diabetes Prediction Workflow

6 Result and Performance Graph

Accuracy Graph

The line of Training Accuracy and Testing Accuracy shakes so much in between the epochs, it is pretty shaky. Two accuracies also intersect 0.65 at different times, namely, at epochs 6 and 21, events which correspond to, respectively, relatively high model performance. But such peaks don't last, and are soon succeeded by dramatic drops that can plummet to as low as 0.30. This kind of variability indicates that the model is overfitting some of the epochs or otherwise making unusual responses to the fluctuations in the data. Such behaviour is generally a sign of bad or wrong choice of regularisation parameters, and as such is unstable.

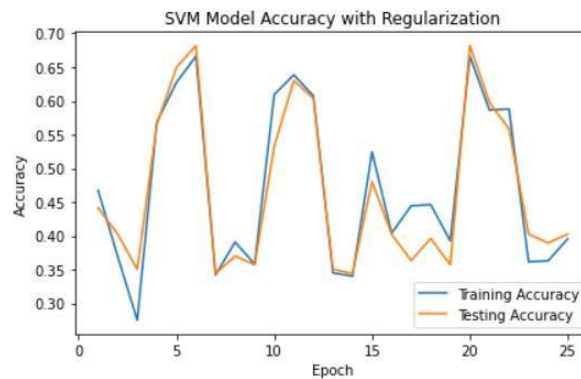


Figure 4: SVM Model Accuracy with Regularisation

Loss Graph

The Training Loss and Testing Loss curves aren't so crazy, but they still have Godzilla jumps. And it looks like a valley around epoch 5 where the testing loss momentarily rises above 1.0 before decreasing again. This bump might have indicated some sort of unstable early training, maybe as a result of initialization or some sudden change in the performance of the model. The two loss curves generally follow one another quite closely, so the model isn't massively overfitting. But log loss numbers are awfully good at meandering around in the 0.7 to 0.9 range, i.e. moderate rates of misclassification, with no clear tendency to converge or improve over time.

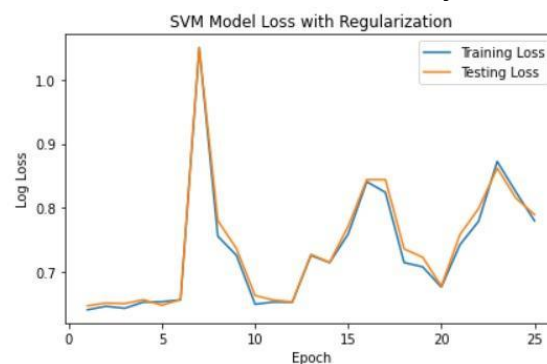


Figure 4 SVM Model Loss with Regularisation

7 Conclusion

The work on Support Vector Machines (SVM) for diabetes prediction shows a remarkable leap in early diagnosis potential. The SVM model accurately separates people by very important factors like glucose levels and BMI, with accuracy rates greater than 77%. It means that SVM is a good classifier to distinguish diabetic patients, which leads to early medical treatment. In future work, deep learning techniques could be used to enhance predictive accuracy even more and to generalise the model to a wider range of healthcare environments.



References

1. J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *N. Engl. J. Med.*, vol. 376, no. 26, pp. 2507-2509, 2017.
2. A. Sarwar and S. Sharma, "Machine learning-based approaches for predicting diabetes," *Int. J. Adv. Res. Comput. Sci.*, vol. 10, no. 3, pp. 92-98, 2019
3. J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease, diabetes, and cancer prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43-48, 2011
4. A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," in *Comput. Sci. and Its Appl.*, 2011, pp. 272-283.
5. S. Joshi and P. Dandannavar, "Machine learning-based diabetic prediction system," *J. Comput. Appl.*, vol. 148, no. 8, pp. 15-19, 2016.
6. P. Kandhasamy and A. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, pp. 45-51, 2015.
7. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104-116, 2017.
8. D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578-1585, 2018.
9. B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102-8108, 2010.