International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Proposed E-SVM Framework for Early Diagnosis of Type 2 Diabetes Mellitus Prediction

Raja S[1], Dr. Nagarajan. L[2]

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: Diabetes is a widely observed metabolic condition distinguished by heightened amounts of glucose in the bloodstream. The identification of this illness in its early stages presents difficulties owing to its intricate reliance on multiple elements. Type 2 diabetes (T2D) is a chronic metabolic disorder that has a substantial impact on a considerable proportion of the global population. The development of crucial decision support systems is necessary in order to provide assistance to medical practitioners during the diagnostic process. The application of many machine learning (ML) algorithms in the prediction of this particular disease has attracted considerable attention, particularly due to its potential for early identification and effective intervention. The present study aims to create a predictive model with the objective of attaining a notable level of classification accuracy in the context of type 2 diabetes. The present study utilizes the Enhanced Support Vector Machine (ESVM) algorithm to predict and screen for diabetes. The dataset used in this study was obtained from Kaggle and consisted of 768 patients, both with and without diabetes, belonging to the Pima Indian population. The dataset under consideration consists of data from 768 patients, encompassing eight primary attributes and a goal column indicating the outcome as either "Positive" or "Negative." The experiment was conducted using the Python programming language, and the results of the demonstration indicate that the utilization of a ML model yields enhanced efficiency in predicting diabetes.

Keywords- *Machine Learning, Kaggle, Positive, Negative, ESVM.*

1. Introduction

Data mining is a highly effective method for extracting valuable insights from datasets that encompass vast amounts of embedded data. The utilization of data mining techniques holds significant potential for effective application within hospital settings, particularly due to the substantial amount of data available. Hospital datasets sometimes need to be clustered and subsequently classified before they can be studied meaningfully. Statistical parameters have been extracted from the dataset using soft computing techniques like pattern recognition (PR) and ML. The World Health Organization (WHO) estimates that around 422 million people around the world have both Type 1 and Type 2 diabetes.

Diabetes Mellitus (DM) encompasses three primary classifications: (i) diabetes mellitus or DM, (ii) insulin resistance, and (iii) gestational diabetes, which is typically observed in pregnant women. Diabetes mellitus (DM) is commonly attributed to elevated levels of blood glucose and is a prevalent condition that affects persons with imbalances in their blood glucose levels, particularly pregnant women. Previous studies have indicated that pregnant women diagnosed with diabetes exhibit a higher likelihood of giving birth to infants with congenital anomalies compared to their non-diabetic counterparts.

Moreover, it is worth noting that in the majority of nations, the prevalence of diabetes has undergone a transformation due to alterations in dietary patterns, the presence of aging populations, diminished levels of physical activity, and the adoption of other detrimental lifestyle behaviors [1]. In general terms, there are two types of diabetes. Insulin-producing cells in people with type 1 diabetes are destroyed, leading to a shortage of insulin. Type 2 diabetes is a disease that happens when the body doesn't make enough insulin or doesn't use it well enough. This happens most often when the body's ability to control blood sugar levels is hampered, causing a relative insulin shortage.

According to the survey conducted by the National Diabetes and Diabetic Retinopathy, the prevalence of diabetes in India is reported to be 11.8% [2]. Based on the available data from the National Health and Nutrition Examination Survey, it is evident that the mortality rate resulting from diabetes surpasses the combined mortality rates attributed to HIV/AIDS, malaria, and tuberculosis. According to a projection made by the World Health Organization, diabetes is anticipated to rank as the seventh most prevalent cause of mortality by the year 2030 [3]. According to the International Diabetes

1Research Scholar PG & Research Department of Computer Science Adaikalamatha College Vallam,Thanjavur Affiliated to Bharathidasan UniversityTiruchirappalli, Tamilnadu, India
Mail.id-sk.rajamecse@gmail.com
[2]Assistant Professor PG & Research Department of Computer Science Adaikalamatha College Vallam,Thanjavur Affiliated to Bharathidasan University Tiruchirappalli, Tamilnadu, India
Mail.id-mcadirector@gmail.com

Federation [4], it is projected that around 482 million individuals worldwide would be impacted by prediabetes by the year 2040, with an estimated 642 million individuals being afflicted by diabetes. The presented results indicate that type 2 diabetes does not exhibit discriminatory tendencies. The phenomenon in question exerts influence on individuals from many social strata, communities, nations, and continents. Unless efficient diagnostic techniques are employed, this issue will continue to be widespread. The aforementioned matter will exert pressure on healthcare systems, provide challenges to economies, and primarily impact the quality of life.

Both patients and doctors seem to like the recommended solution. A precise early identification system for diabetes may be worth exploring for countries with a high diabetes prevalence. According to reference [5], the accuracy of estimating the severity of an underlying medical problem improves with the level of precision in the prognosis. Those who suffer from diabetes would benefit greatly from having their illness diagnosed and treated as soon as possible. The implementation of computer-based systems for early detection offers advantages to both patients and medical personnel. By replacing the usual manual analysis of data, these systems can enhance efficiency, reduce costs, save time, and minimize the potential for human errors. The user did not provide any text to rewrite.

This article's subsequent sections are organized as follows: In Section 2, we present a synopsis of the relevant literature, while in Section 3, we go into depth on the data and methods used to develop this strategy. Results and interpretation of the data are presented in Section 4, and the paper is wrapped up in Section 5.

2. Literature Review

For the purpose of developing a highly accurate predictive model, the existing body of literature demonstrates the use of a wide variety of data mining and machine learning techniques, such as neural networks (NN), tree-based algorithms, rule-based fuzzy classifiers, linear classifiers, and hybrid models. The Artificial Neural Network (ANN) emerged as the predominant classifier, as evidenced by its utilization in 11 out of the 20 previous investigations, wherein various varieties of ANN were employed in the experimental procedures. A majority of the initial experiments yield prediction accuracies falling within the range of 80% to 90%.

Wu et al [6] employed a novel hybrid predictive model and achieved a remarkable accuracy of 95.42% in their predictions, which stands as the highest reported in the existing literature. It is probable that medical datasets often exhibit a substantial number of missing data values. The suggested solutions put forth by the authors were thoroughly examined, as they had a significant impact on the overall correctness of the categorization..

Ensan, et al in [7] introduced Fuzzy Clustering (FACT) method in one of their earlier studies. This method suggests that the determination of the appropriate number of clusters should be based on the density of the data. However, it should be noted that the suggested approach demonstrates insensitivity towards the starting number of clusters, a common practice where the initial cluster values are typically adjusted to be lower than the threshold range of clusters. The approach employed by the researchers involves the creation of many clusters, with a specific emphasis on identifying and isolating outliers. The authors of the study conducted an experimental analysis to demonstrate that the heuristic algorithm they devised outperforms the traditional K-means computation in terms of performance.

Purnami et al in [8] introduced the Support Vector Machine (SVM) classification algorithm for detecting Data Mining (DM). In this study, the authors put forth an enhanced iteration of the Support Vector Machine (SVM) algorithm, specifically referred to as Smooth SVM (SSVM) and Multiple Knot Spline SSVM (MKS-SSVM). On the PIMA dataset, where both algorithms were tested, the researchers found that the MKS-SSVM technique provided more accurate results.

Aishwarya et al in [9] presented a method based on fuzzy logic to examine the process of selecting choices. The model found an occurrence link and an accuracy relationship between symptoms and diseases. The symptom's repeatability is confirmed by the occurrence relationship, while the disease's presence is depicted by the confirmability relationship. Additionally, the researchers introduced the concept of Fuzzy Logic, employing the notions of minimum and maximum relationships. They conducted an empirical study using a dataset consisting of 40 patients to examine the fuzzy relationship in the context of examining Diabetes Mellitus (DM).

Kumari and Chitra in [10] have conducted a study that establishes correlations between several ML and data mining approaches in the context of predicting diabetes mellitus (DM). The authors employed SVM in their study. Interpretable SVMs have a strong potential for effectively predicting the occurrence of diabetes, as shown by the results of exploratory study performed on an authentic dataset relevant to diabetes. The preliminary findings of this inquiry reveal that the SVM model achieved an accuracy rate of 79.00%. It has been found that using feature subset selection within SVM classifiers improves their performance.

Kumar et al., in [11] suggested the utilization of Genetic Algorithm (GA) in conjunction with a SVM classifier to identify an appropriate subclass of components across many datasets, with the aim of enhancing the accuracy of characterization. The classifier based on genetic algorithms appears to enhance the estimation of boundaries for support vector machines by selecting the optimal collection of characteristics. The evaluation of results using the established SVM approach demonstrated that the proposed strategy achieved an accuracy rate of 83.00%.

Jhaldiyal and Mishra in [12] employed Principal Component Analysis (PCA) and SVM to forecast the occurrence of diabetes in patients. The study's investigational findings indicate that the previous stage can be improved, as it achieved an accuracy in classification approximately 93.66%.

Vispute et al. in [13] examined the potential association between order procedures and the likelihood of developing diabetes mellitus (DM). The study examined four machine learning processes, specifically Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR) and Artificial Neural Networks (ANN). Within this application, they employed the ROC curve approach to predicting in the domain of data mining. The user enters data into the program, and the program returns results that include both actual and predicted statistics. Ultimately, the conducted studies provided empirical evidence that Random Forest (RF) exhibits a high level of accuracy.

Lingaraj et al., in [14] conducted an experiment on detecting DM using the WEKA mining tool, employing the Navie Bayes classifier. The study utilized data that was gathered from hospitals in India, consisting of a total of 1865 cases. These instances included variables related to blood tests and urine tests. After running an experiment using 10-fold cross-validation, the researchers compared the results. In spite of the authors' use of 10-fold validation, their computer model achieves only an accuracy of 84.89%. This shows only a very modest performance, and further study is clearly required.

Perveena et al., in [15] (2016) utilized the DT J48 algorithm to detect DM, focusing on the identification of risk factors. The authors of the study showcased the superior efficiency of Adaboost in comparison to both bagging and DT J48 within their framework. Due to the lack of basis learners within the ensemble structure, this study is limited in terms of several performance indicators. Consequently, there exists a research vacuum that can be addressed by incorporating base learners into a classifier ensemble.

Cui et al., in [16] proposed a hybrid estimation model that incorporates principal component analysis (PCA) into the original dataset. Subsequently, the C4.5 algorithm was employed to develop the classifier model. The accuracy of their work's classification was approximately 89.0%, a value that can be improved with the implementation of appropriate feature selection techniques.

Haritha et al., in [17] examined an alternative data mining methodology for detecting DM. This strategy involved the integration of classification approaches with the use of association based rule mining techniques. The classification execution technique was evaluated by employing the KNN classifier. The authors initially employed a set of ten features to conduct the K-nearest neighbors (KNN) analysis, resulting in an accuracy rate of 61.9%. Subsequently, the authors employed Particle Swarm Optimization (PSO) techniques to choose six features for training, resulting in a notable enhancement of the accuracy to 88.5%.

Rashid and Abdullah in [18] examined the detection of Type-1 and Type-2 diabetes mellitus (DM) by employing firefly and cuckoo search algorithms to pick variables from the Indian PIMA dataset. The authors utilized the firefly algorithm to pick the ideal features, and subsequently employed the K-nearest neighbors (KNN) classifier for predicting DM. In contrast, the optimal features obtained by the cuckoo search strategy were utilized in conjunction with the Fuzzy-KNN classifier. Based on the findings of their empirical investigation, the researchers arrived at the conclusion that the cuckoo-Fuzzy KNN algorithm demonstrated the highest level of classification accuracy. It is also possible to speed up an optimization method's learning process.

The SVM model proposed by Perveen et al., in [18] provides valuable insights into the diagnosis of DM. The authors of this study regarded diabetes mellitus (DM) as a substantial global health issue and discovered that 80% of complications associated with DM can be mitigated when they are detected in the early stages. The present study evaluates a range of data mining and ML algorithms for the purpose of forecasting data mining. For instance, the researchers put forth a Support Vector Machine (SVM) model that incorporates an additional module to transform the "discovery" model of an SVM into a defensible representation. This unique approach offers the opportunity to employ Support Vector Machine (SVM) classification with a notable degree of accuracy. However, there is room for improvement in enhancing performance across all sample scenarios.

Sivakumar et al., in [19] aimed to investigate the efficacy of several artificial intelligence (AI) models in identifying DM infection by analyzing patterns in the available data. The objective was to provide timely therapy to patients based on these findings. The AI classifiers, specifically the NB calculation and RF algorithm, achieved precision rates of 76.3% and 75.7% respectively when applied to effectively ordered samples.

Srivastava et al., in [20] proposed a method that incorporates the Fuzzy C-Means Clustering (FCM) method in conjunction with Support Vector Machines for the purpose of diabetes prediction. The model achieved an accuracy rate of 84.24%. The authors of this work have used a FCM algorithm to effectively handle the elimination of unwanted data. However, they have ensured that the computational trade-off associated with FCM has not been compromised. The inclusion of an automatic approach to determine the appropriate number of cluster requirements is necessary for the implementation of this algorithm.

3. Methods and Materials

3.1 Dataset

Our analysis centers on the PIMA dataset obtained from the official website of the University of California, Irvine (UCI) (https://www.kaggle.com/uciml/pima-indians-diabetes-database). This dataset has been extensively examined in prior research. As depicted in Table 1, the dataset consists of 768 samples that possess specific features. In order to aid in the process of imputing missing values from the dataset, we employed correlation as a pre-processing step. The correlation coefficient can be utilized to estimate the existing causal relationship between the provided data. The pre-processed attributes will subsequently be employed in subsequent stages.

Table 1: Dataset Description

Attributes	Description
No of Samples	768
No of Features	8
Output Classes	2
No of total features	9
No of Missing attributes	Null
No of Noisy attribute	Null

3.2 Proposed ESVM

The achievement of optimal classification performance is contingent upon the utilization of effective data preparation and pre-processing techniques. The algorithm that has been developed addresses the research gap that was highlighted in the literature review. The performance of a ML model can be significantly enhanced by employing a meaningful technique to address missing values in attributes.

This study examines the utilization of the Extended Support Vector Machine (ESVM) classifier for the purpose of analyzing the PIMA Indians Diabetes Database. The general architecture of our suggested innovative technique is seen in Figure 1.

In order to assess the accuracy and reliability of a predictive model for certain factors, it is imperative to subject it to rigorous testing using both known and unknown data. In the process of developing the prediction system, our effort involves categorizing the supplied dataset using specific heuristic principles. The proposed approach employs the Elastic Support Vector Machine (ESVM) technique. The training procedure will be conducted on the given training dataset, utilizing the computation efficiency of the ESVM. To assess the effectiveness of the suggested method, it was additionally examined for its ability to predict DM on test data. In order to optimize the user experience, these functionalities have been included into a User Interface (UI) in the form of a sequential application flow.

In a concise manner, the dataset was loaded and afterwards subjected to pre-processing techniques in order to remove any instances of null values. Subsequently, the pre-processed data was shown through confusion matrix and histogram plots before implementing the SVM algorithm to produce accurate predictions. Figure 1 illustrates the comprehensive depiction of the suggested concept.

SVM is a widely utilized supervised learning method that is regularly employed for regression and classification analyses. The algorithm operates by encoding each data point as a coordinate in an n-dimensional space, with the feature value serving as the coordinates. Figure 1 depicts the binary classification framework, whereby C1 and C2 represent distinct classes. SVM employ the optimal margin for the purpose of classifying the output.

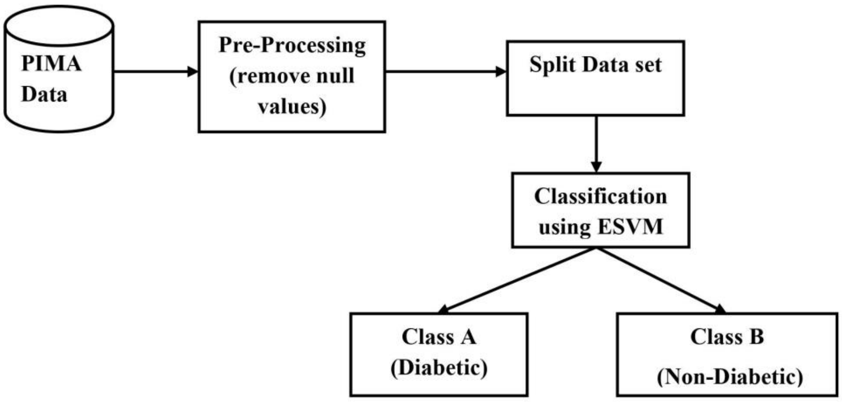

Figure 1: Block Diagram of Proposed ESVM

Consequently, the SVM approach was employed to identify an appropriate classifier by employing the Gaussian function on a collection of training datasets.

A. Instance Learning

The primary objective of manifold learning is to develop a model capable of discerning a collection of predetermined good and bad instances. Each unit comprises numerous occurrences. The acceptance of a bit as positive is contingent upon the presence of at least one positive case, whereas equally, a bit is deemed negative if all cases exhibit negativity.

Given a set of bit parameters P_1, P_2, …, P_i, it is observed that there are m occurrences for each bit P_i, where the value of P_i is less than 1 and greater than m. Additionally, each bit is associated with a corresponding label. Each bit parameter P_i is associated with a label X_i, which falls within the range of (-1,1), while maintaining the absence of simplification. If the value assigned to a bit parameter is negative, all instances within the bit will also have negative labels, as indicated by the following equation:

$$X = \sum_{i=1}^{m} \frac{X_i + 1}{2} \quad (1)$$

Here the value of $X \geq 1$.

B. Feature Selection

The initial recommendation of the twofold classification technique of Support Vector Machines (SVM) was based on its ability to effectively handle complex nonlinear boundary models. However, this method of calculating parameters incurs significant processing costs. Therefore, we offer an improved multi-instance technique that relies on the Support Vector Machine (SVM) algorithm. The instance feature selection method proposed by Wang et al. (2019) is similar to this approach. Small sample sizes, nonlinearity, and high-dimensional design perception are all areas where our proposed approach shines. The fundamental objective is to discover a discriminating function that can assess the instance parameters in light of the given imperative.

The assignment of a label to a bit is determined by the superior instance within the bit's structure. In equation (1), it is observed that the bit contains just one negative tag. Consequently, the SVM algorithm facilitates the ordering of values by establishing many hyper planes in a multidimensional space. This allows for the segregation of cases into different class labels. Support Vector Machines (SVM) possess the capability to handle an extensive number of features without any limitations. By utilizing Support Vector Machines (SVM) to train a model, we are able to effectively define these properties in a suitable manner. To mitigate the complexities associated with the traditional Support Vector Machine (SVM), we offer an improved version of SVM that involves the modification of the parameters E, epsilon (ε), and bit. In general, Support Vector Machines (SVM) are capable of identifying an edge that effectively separates positive and negative models.

The cost associated with precisely aligning the characterization is denoted as E. The parameter E serves as a trade-off between the misclassification of training models and the generalization of the decision boundary. The selection of a low value for E results in a smooth decision surface, whereas a high value for E aims to precisely rank all training instances by allowing the model to choose more appropriate examples as support vectors.

The selection of the parameter ε is necessary for the estimation of the inhumane loss function. The parameter ε has an impact on the efficacy of the Support Vector Machine's response and it also effects the number of support vectors. Consequently, both the complexity and the stability of the system are expected to be affected by its value. The option "bit" is commonly known as the penalty parameter, which allows for the inclusion of errors. The initial design of an advanced edge classifier necessitates the use of separable class distributions and does not tolerate any errors in training data points. The penalty parameter serves to carefully and fluently constrain the attention towards an unsuitable aspect of the decision boundary.

Support Vector Machines (SVM) possess the capability to effectively handle vast feature spaces and effectively process large-scale information. The issue of overfitting can be mitigated by employing a refined boundary method. In the context of Support Vector Machines (SVM), the training process is typically straightforward, but the selection of an appropriate kernel function might be challenging. The Enhanced Support Vector Machine (SVM) employs a unique parameter configuration in order to avoid misclassifying patterns.

C. Algorithm 1: Proposed ESVM

```
Input: Data Set S ($E_i$ and $X_i$ optimal Parameters for ESVM)
Output: Classification
Step 1: Initialize parameter i, E for all the classes
Step 2: Find parameter ε and bit based on SVM
Step 3: find new parameter $E_i = \varepsilon^j a_i$ for all the instances
Step 4: for each positive parameter $X_i$ do
        If $\frac{\sum_{i \in 1}(X_i+1)}{2} == 0$ then
            Calculate I’=argmax $E_i$
            Update $X_i' = 1$
Step 5: End if
Step 6: While The prior round's label for the instance has changed do
Step 7: End While
Step 8: Print ($X_i$, updated bit)
```

4. Result and Discussion

The dataset used in this study is the PIMA Indian diabetes dataset, which was obtained from the UCI repository (uci.edu). The proposed analytical approach was developed using Python 3.6.4, along with several libraries such as Keras, TensorFlow, Scikit-Learn, Pandas, Matplotlib, and other essential libraries. The ESVM algorithms were employed to assist in the identification of cases of DM. The findings indicate that the ESVM model, as proposed, achieves an accuracy rate of around 98.45%. The outcomes of SVM and ESVM are compared and shown here.

Table 2 displays the confusion matrix, which represents the alignment of true and expected conditions inside a single grid. As evidenced, the achievable result of a natural job can be classified into one of four distinct groups.

Table 2: Confusion Matrix

	Predicted Values	
Actual Value	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Positive data can be seen as representing accurate and precise information, whilst negative data may indicate inaccuracies within the dataset.

A. Accuracy (Acc)

Accuracy refers to the measure of correctly predicted instances of diabetes classification, either for an independent test set or through the implementation of cross-validation techniques. It is a fundamental metric used to assess the effectiveness and reliability of predictive models,

$$Accuracy\ (Acc) = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

B. Sensitivity (Sen)

Sensitivity refers to the proportion of genuine positive cases correctly identified by a diagnostic test or screening tool. In the problem domain being examined, the focus is on determining the degree to which data pertaining to individuals with diabetes is reliably identified as indicative of the condition, which is defined as,

$$Sensitivity\ (Sen) = \frac{TP}{TP+FN} \quad (3)$$

C. Specificity (Spe)

In a general sense, sensitivity is indicative of the test's ability to accurately predict the appropriate classification. The concept of specificity refers to the genuine negative rate. This study examines the prevalence of misdiagnosis among individuals with diabetes, specifically focusing on the extent to which individuals with diabetes are wrongly identified as having the condition.

$$Specificity\ (Spe) = \frac{TN}{TN+FP} \quad (4)$$

In a general sense, the concept of specificity pertains to the degree to which a given test accurately predicts the occurrence of a particular classification. Table 3 displays the confusion matrix, whereas Table 4 shows the evaluation metrics obtained using our proposed approach. Figure 2 shows the proposed ESVM's confusion matrix.

Table 3: Values obtained in Confusion Matrix

	Predicted Values	
Actual Value	177	7
	4	523

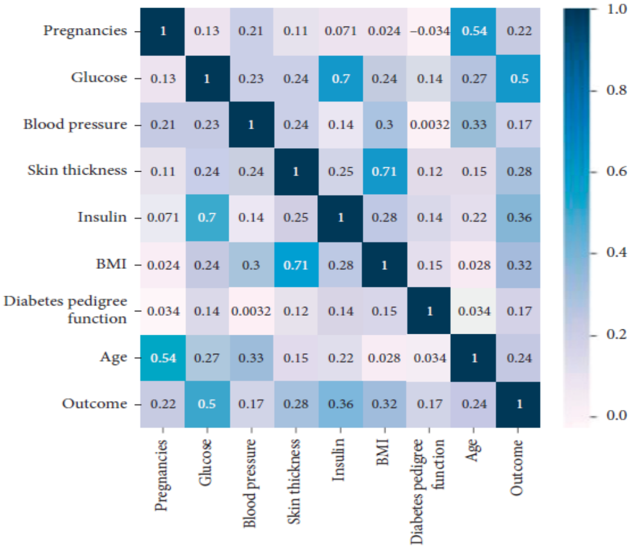

Figure 2: Confusion Matrix of ESVM

Table 4: Performance evaluation Metrics of Proposed work

Performance Measures	Values
Number of features used	5
Sensitivity	98.4
Specificity	94.6
Accuracy	98.5

Figure 3 presents a visual representation of the comparative analysis conducted to evaluate the accuracy, sensitivity, and specificity values of the suggested methodology in comparison to existing methodologies, as presented in Table 5. Based on the comparing data, it is evident that the ESVM-DNN approach demonstrated greater results for all performance metrics.

The retrospective examination of our research findings is now possible. Initially, the minimum amount of attributes selected by the Extended Support Vector Machine (ESVM) algorithm. Furthermore, the highest achieved classification accuracy using the Extended Support Vector Machine (ESVM) method is 98.5%. Among the total of 768 occurrences, the ESVM technique, as proposed, successfully recognized 711 samples as correctly classified, whereas 177 samples were identified as outliers. The PIMA diabetes dataset considers pregnancy and age as crucial variables. The ESVM technique shown superior performance in key parameters, including accuracy, sensitivity, and specificity evaluations, when compared to existing methodologies.

Table 5: Performance comparison of proposed Vs Existing

Measure	SVM	Proposed ESVM
Sensitivity (%) [12]	82.3	98.4
Specificity (%) [12]	89.7	94.6
Accuracy (%) [12]	93.6	98.5

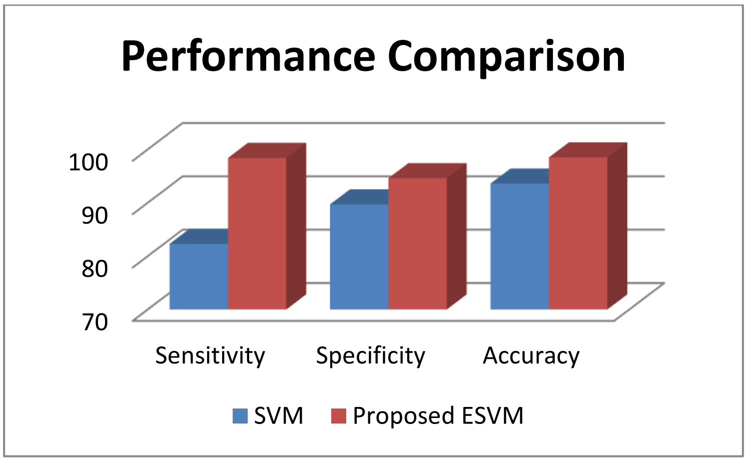

Figure 3: Performance comparison of Proposed Vs Existing

5. Conclusion and Future Work

The application of intelligent machine systems, facilitated by advancements in technology and enhanced computational capabilities, offers a more efficient alternative to traditional manual analysis in the context of medical decision support. The current solutions involve categorization models that demonstrate a notable level of imprecision. However, these models fail to incorporate crucial data preparation procedures that have the potential to enhance the total accuracy attained. The objective of this study is to develop a forecasting model that can accurately and efficiently estimate the occurrence of Type 2 Diabetes Mellitus (T2DM). As previously said, a significant segment of the global populace has a heightened susceptibility to T2DM, hence amplifying the potential for the creation of novel viruses and infectious ailments.

In this study, we want to evaluate the effectiveness of the ESVM classifiers on the PIMA Indian database with the purpose of demonstrating the potential of machine learning algorithms in reducing risk factors and enhancing the accuracy of T2DM screening. The outcomes obtained from our methodology applied to the PIMA Indian dataset exhibit superior performance compared to the SVM model used to a comparable dataset. Furthermore, it is worth considering the potential extension of this research to incorporate novel optimized feature selection techniques prior to model training for classification purposes. Achieving significant advancements in uncovering latent knowledge from valuable datasets stored in crucial databases necessitates the utilization of continuous investigations and cutting-edge algorithms.

References

[1] J. C. Pickup, “Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes,” *Diabetes Care*, vol. 27, no. 3, pp. 813–823, 2004.

[2] N. C. Sharma, Government survey found 11.8% prevalence of diabetes in India, 2019, https://www.livemint.com/science/ health/government-survey-found-11-8-prevalence-of-diabetesin- india-11570702665713.html.

[3] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, “Global prevalence of diabetes- estimates for the year 2000 and projections for 2030,” Diabetes Care, vol. 27, no. 5, 2004.

[4] World Health Organization, Global Report on Diabetes, WHO, Geneva, Switzerland, 2016, https://www.who.int/ publications/i/item/9789241565257.

[5] International Diabetes Federation, IDF SEA Members., 2020, https://idf.org/our-network/regions-members/south-east-asia/members/94-india.html.

[6] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100–107. doi:10.1016/j.imu.2017.12.006.

[7] Ensan, F., Yaghmaee, M. H., & Bagheri, E. (2006). FACT: A new Fuzzy Adaptive Clustering Technique. The 11th IEEE Symposium on Computers and Communications. doi: doi:10.1109/ISCC.2006.73.

[8] Purnami, S. W., Embong, A., Zain, J. M., & Rahayu, S. P. (2009). A New Smooth Support Vector Machine and Its Applications in Diabetes

Disease Diagnosis. Journal of Computational Science, 5(12), 1003–1008. doi:10.3844/jcssp.2009.1003.1008.

[9] Aishwarya, R., Gayathri, P., & Jaisankar, N. (2013). A method for classification using machine learning technique for diabetes. IACSIT International Journal of Engineering and Technology, 5, 2903–2908.

[10] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797–1801.

[11] Kumar, G. R., Ramachandra, G. A., & Nagamani, K. (2014). An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets. International Journal of Advanced Research in Computer Science and Software Engineering, 4(2), 272–277.

[12] Jhaldiyal, T., & Mishra, P. K. (2014). Analysis and Prediction of Diabetes Mellitus Using PCA, REP and SVM. International Journal of Engineering and Technical Research, 2(8), 164–166.

[13] Vispute, N. J., Sahu, D. K., & Rajput, A. (2015, December). A Survey on naive Bayes Algorithm for Diabetes Data Set Problems. International Journal for Research in Applied Science and Engineering Technology, 3(12).

[14] Lingaraj, H., Devadass, R., Gopi, V., & Palanisamy, K. (2015). Prediction of Diabetes Mellitus using Data Mining Techniques: A Review. Journal of Bioinformatics & Cheminformatics.

[15] Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2019). Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. IEEE Access. IEEE, 7, 1365–1375. doi:10.1109/ACCESS.2018.2884249.

[16] Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. [PubMed]. Computer Methods and Programs in Biomedicine, 166, 123–135. doi:10.1016/j.cmpb.2018.10.012.

[17] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo Search Algorithms. International Journal of Applied Engineering Research: IJAER, 13(2), 896–907.

[18] Rashid, T. A., & Abdullah, S. M. (2018). A hybrid of artificial bee colony, genetic algorithm, and neural network for diabetic mellitus diagnosing. ARO-The Scientific Journal of Koya University, 6(1), 55–64. doi:10.14500/ aro.10368.

[19] Sivakumar, S., Venkataraman, S., & Bwatiramba, A. (2020). Classification Algorithm in Predicting the Diabetes in Early Stages. Journal of Computational Science, 16(10), 1417–1422. doi:10.3844/jcssp.2020.1417.1422

[20] Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A Rule-Based Monitoring System for Accurate Prediction of Diabetes: Monitoring System for Diabetes. International Journal of E-Health and Medical Communications, 11(3), 32–53.