Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM

Ahmed Hamza Osman
Department of Information System, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

Hani Moetque Aljahdali
Department of Information System, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

***Abstract*—Now-a-days, diabetes disease is considered one of the key reasons of death among the people in the world. The availability of extensive medical information leads to the search for proper tools to support physicians to diagnose diabetes disease accurately. This research aimed at improving the diagnostic accuracy and reducing diagnostic miss-classification based on the extracted significant diabetes features. Feature selection is critical to the superiority of classifiers founded through knowledge discovery approaches, thereby solving the classification problems relating to diabetes patients. This study proposed an integration approach between the SVM technique and K-means clustering algorithms to diagnose diabetes disease. Experimental results achieved high accuracy for differentiating the hidden patterns of the Diabetic and Non-diabetic patients compared with the modern diagnosis methods in term of the performance measure. The *T*-test statistical method obtained significant improvement results based on K-SVM technique when tested on the UCI Pima Indian standard dataset.**

***Keywords*—K-means Clustering; Diabetes Patients; SVM; Diagnosis; Accuracy**

I. INTRODUCTION

Diabetes disease is an incessant malady that happens when the pancreas does not create sufficient insulin or if the body cannot viably utilizes the insulin, it makes. Insulin is a hormone that controls blood sugar. Hyperglycemia, or raised blood sugar, is a typical impact of unrestrained diabetes and after sometimes prompts actual harm to vast numbers of the body's frameworks, particularly the arteries and veins. In 2004, an expected 3.4 million users passed on from results of fasting high blood sugar[1]. Some studies that have been conducted recently presented common diseases that are frequently misdiagnosis therein; as the number of dead because of medical errors each year to nearly 98,000 people. The therapeutic analysis is viewed as an essential yet confused errand that should be executed precisely and proficiently. The mechanization of this system would be to a significant degree beneficial. Unfortunately, all specialists don't have aptitude in each sub Specialty and also there is a deficiency of asset people at specific spots. Subsequently, a program medicinal determination system would presumably be exceedingly valuable[2]. The objective of this study is to develop a hybrid technique based on Support Vector Machine algorithm and Two-step clustering method for diabetes diagnosis. The proposed method seeks to reduce the ratio of misdiagnosis of diabetes and increase the ratio of accuracy for diagnosis.

Nanda et al.[3] proposed a Classification of Gestational Diabetes Mellitus's (GDM) framework for biological and maternal features at (11 to 13) weeks gestation. The benefit is that a combining of maternal characteristics and biomarkers for GDM can provide First-trimester screening. Alssema et al.[4] updated a risk diagnosis survey using type2 diabetes screening detecting approach. They have considered other predictors for diabetes detection, but the drawback is that although particularly the case for small data needs an external validation before applying a model. Bennetts C.J et al.[5] intended for exploited a relationship of the diabetes dataset for classification capability. Like many researchers, Bennetts C.J et al.[5] tried to employ the data mining approaches in the biological organization for discovering new knowledge which assists to help a medical doctor for accurate diagnosis. A brief overview is provided by Tomar et al.[6] of these methods and their advantages and drawback. Diabetes treatment based on predictive analysis was presented by Aljumah et al.[7] using regression classification technique. They developed a mining tool called Oracle Data Miner (ODM) for treating modes diabetes prediction and SVM for results analysis. Kalaiselvi and Nasira[8] proposed a combination of PSO and SVM methods for to test the relationship of diabetes and heart disease. Their proposed method tried to extract the association factors disease based on categorical features which are the main benefit of the PSO-SVM method. Saudek et al.[9] developed a diabetes diagnoses measures of screening for finding the patients and clinicians rapidly. The determined criteria named HbA1care recognized for screening and now described as IFG.

Up to now, several studies described that have attentive on therapeutic diagnosis. These researchers had introduced various methods to the assumed challenges and obtained high prediction precisions, of 77% or greater, using the UCI Pima Indian dataset [10]. Empirical results proved precise prediction accuracy of 77% with logistic-regression derived discriminated function. Breault and colleagues [11] proposed a regression tree (RT) as a classifier method applied on data of 15,902 diabetes persons. The results select a greatest significant variable related to inadequate glycemic control >9.5. There are many models and methods used by scientists to examine and diagnosis diabetes. One of these models is a support vector

machine. The SVM model generation is a group of the relevant supervised-learning technique used in health diagnosis for regression and classification [12] and [13]. It is a standard method based on guaranteed risk limits of statistical learning theory e.g. the called structural risk minimization principles. Athanasios et al. [35] Reviewed the methodology that was proposed by Dalakleidi et al.[36]. Thier review study discussed a combination method between the K-NN classifiers and genetic algorithm to define the critical risk factors that are robustly related to the occurrence of non-fatal and fatal Cardio-vascular Disease (CVD) in with Type 2 Diabetes patients Mellitus(T2DM). Tao Zheng et al.[37] introduced a T2DM model for identifying subjects using machine learning and feature engineering. The model contrasted and evaluated the identification performance using different machine learning such as Naïve Bayes, Decision Tree, k-NN, Random Forest, SVM. The performance model used the Area Under the Carafe (AUC) evaluation measure and obtained accuracy average with 0.98.

Based on the previously mentioned, issues identified with diabetes are numerous and entirely exorbitant. It is a serious malady cause, if not treated legitimately and immediately, it could prompt significant confusions, may be the demise of the patient. These made diabetes one of the principle needs in therapeutic knowledge study. However, the country has not been utilized the strength of computer technology to reduce the risk of diabetes yet. With the rise of the new knowledge, scientists have discovered various kinds of new technologies that the developed research could use to solve this problem. One of the main promising technologies today is Knowledge Discovery. It is capable of predicting the risk level of a patient with significantly higher accuracy by extracting hidden patterns from historical medical records. This reason will help us to give timely treatment for patients by diagnosing disease early before it goes to a critical stage.

Rest of this research is organized as Sections 2 explain the materials and method. Section 3 presents the experiments of the suggested method. Section 4 discusses the results of the combined technique. Section 5 reported the conclusions of the study.

II. Ease of Use

An operational framework of the hybrid K-SVM method divided into three main work stages; each of these steps consisted of different phases, starting from the data preparation stage and ending with the diagnosis phase as implementation stage. Figure 1 shows the operational framework stages of the introduced technique (SVM-K-mean clustering).

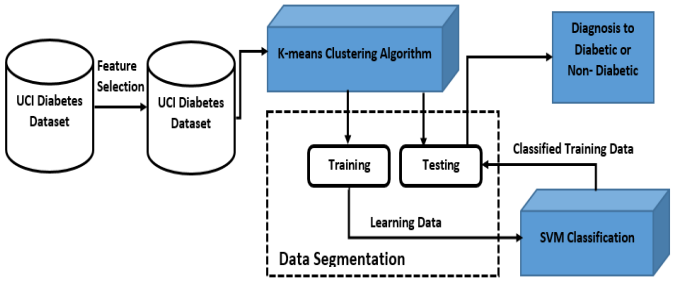


Fig. 1. Operational model of proposed method

A. Phase 1: Preparing and study the dataset

Data Collection: The diabetes dataset called Pima Indian collected from UCI machine repository standard dataset. This dataset used with different fields and research such as [7,12,13 and 14], is a gathering of symptomatic therapeutic reports from 768 records of female patients no less than 21 years of age of Pima Indian legacy, a populace living close Phoenix, Arizona, USA. The binary target variable takes (0 or 1) values, while 0 implies a negative test for diabetes, and 1 indicates a positive test. There are 500 cases in class 0 and 268 cases in class 1. Fine-tuning parameters further physically assessed the significance of the systematically selected of variables. The variables incorporated into the last determination were those with the best discriminative execution.

TABLE I. The Pima Indian Dataset

No	Feature
1	Number of times pregnant
2	Plasma glucose concentration 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function
8	Age (years)
9	Class variable (0 or 1)

The Pima Indian dataset is reported that there are no missing values; there were some generously included zeros as missing values. 28 cases had a diastolic blood pressure of 0, five patients had a sugar of 0, 11 more had a mass body record of 0, 140 others had serum insulin levels of 0, and 192 others had a skin fold thickness readings of 0. After the erasure, there were 460 cases with no missing values.

Feature Selection (FS): It is significant that the data set is pre-processed before mining process is used so that repeated data can be removed or the unstructured data can be counted by transformation of the dataset. Theoretical strategies for selecting proper features differ for a different challenge to another. Employing feature selection is the important step to simplifying the learning part of the mining stages and enhancing the performance without altering the primary structure of data mining methods [15] and [16]. The proposed method used the feature selection algorithm as a preprocessing step to reduce the dataset dimensionality and increased the computational process as well.

Data Segmentation: Because of a significant amount of associated samples datasets and examine the dataset with different data size, the developed method divided the dataset into 50%, 60%, and 70% for training and 50%, 40%, and 30% for testing as a preprocessing step for the proposed approach[17].

B. Phase 2: Diabetes data clustering using K-means algorithm

It is a very useful approach to machine learning for classification of native clusters in a dataset means clustering

algorithm, as groups of data are based on their feature values to K clusters. In classification, the items are assigned to pre-defined classes, whereas in clustering the classes are formed [18]. K-means algorithm is one of the hard clustering approaches; therefore, a data point can belong to just one cluster [19]. The proposed method applied a description of diabetes data using the K-mean method to group the diabetes data based on the Euclidean distance similarity of their features into K clusters [20] and [21]. Before the algorithm begins, K is a positive number initialized early to refer to the number of required clusters [22]. K-means group inspects the diabetes feature of each patient to ensure that the elements data within each group are similar to each other but dissimilar from items in other groups.

The algorithm initiated by selecting an initial set of groups repeatedly updated until no further improvement can be made or until a specified limit is exceeded by the number of iterations [23] and [24]. The developed technique used it to measure the difference between the patient's data (Euclidean distance is used as a measure to define the similarity between data items) 25. See Eq. 1.

$$(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

A Euclidean vector is the position of a point in a Euclidean n-space. Therefore, X (Xn, Xn,. . . Xn) and Y (Y1, Y2,. . . , Yn) are Euclidean vectors21. Figure 2 shows the steps of K-means approach (demonstrates K-mean clustering steps).

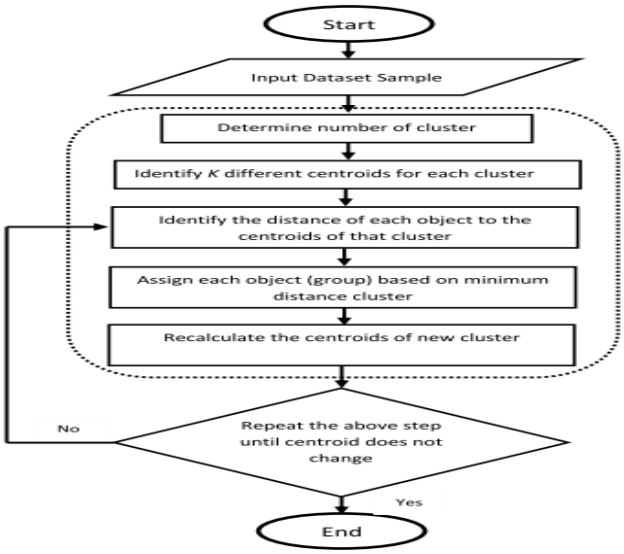


Fig. 2. K-means Clustering Stages

C. Phase 3: Diabetes Prediction and diagnosis using K-SVM

The Support Vector Machine is a promised classifier method widely used because of the significant output that had extracted in different research areas, and cause of their robust assumed, theoretical underpinnings in the theory of statistical learning[26]. SVM classification proceed the hyperplane in distinct classes [14]. Every hyperplane is identified by its path (w), the correct position threshold is (b), the dimensionality input is (xi) and pointed the class. A set of the learning samples is represented using formula 2 and 3.

$$(x1, y1), (x2, y2), \ldots, (xk, yk); \; Xi \in Rd \qquad (2)$$

Where k is the learning data, and d denotes to dataset dimensions number:

$$yi \in ; \; i = 1, 2, \ldots, k$$

The decision equation of the form formula. 2.

$$f(x, w, b) = \text{sgn}\big((w \cdot xi) + b\big), w \in Rd, \quad b \in R \qquad (3)$$

The region between the hyperplane is a margin, which splits two classes; the margins illustrate the prediction of diabetic patients by the SVM classifier. The $\frac{1}{||w||}$ represent the distance between the hyperplane and closed data point. Figure 3 demonstrates the diagnosis of diabetes data using SVM.

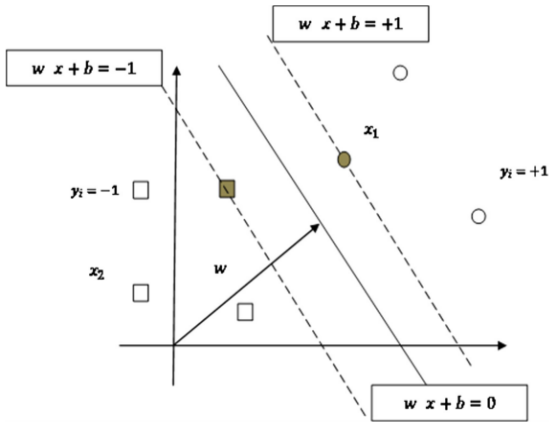


Fig. 3. Classification of Diabetes using SVM

The introduced hybrid method is a technique for diabetes dataset diagnosis by K-mean clustering algorithm and SVM technique and comprises of two sub-techniques: unsupervised learning based on K-means data clustering using features similarity, and classification diabetes using the supervised SVM classifier method. The objective of this study is to produce a diabetes patient's diagnostic process with the assistance of a combined K-means clustering and SVM method for the improving the diagnosis accuracy and to reducing the misdiagnosis error. The proposed method considered the clinical analysis output based on pancreas cell and Laboratory test results of patient blood and urine. The clustering output will then uses as input features for the SVM classifier of the patinas data. The clustering process utilized to group all similar instances in feature data, which enhances and increases the predictability and accuracy of diagnosis using SVM classification method.

III. Experimental Design

The experiments of this research conducted by using the UCI-Pima Indian dataset for the evaluation of the KSVM technique. To assess the performance of classifiers, the K-SVM expriment was ran frequently to let the entire slice in the dataset to take an opportunity as a testing data. To decrease the feature dimensionality a feature selection method is used[15]. This process employed independently before the training of the original dataset. Figure 4 demonstrate the extracted significant features.

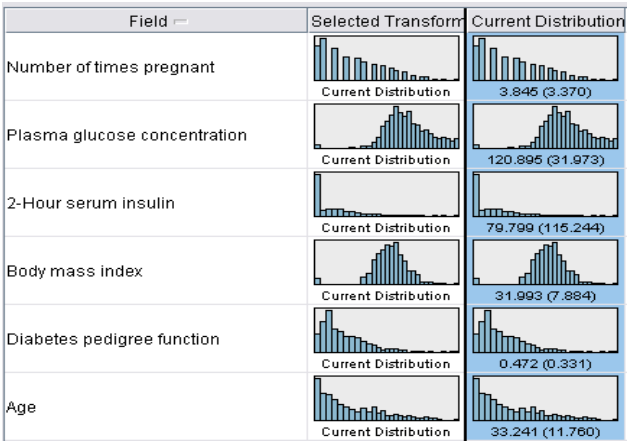


Fig. 4. Diabetes extracted features

In each round, the data was divided into 3 clusters (70%, 60%, and 50%) for learning procedure and (30%, 40%, and 50%) for the testing process[17]. The K-means technique conducted for grouping patients based on related target Diabetic and Non-Diabetic features. Due to the need for overall reduction in distance between cluster centroids and cluster members, the K-means formulated by using an optimization problem13.

$$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_{k=1}^{k} \sum_{i \epsilon S_k} \left\| x^i - \mu_k \right\|^2 \quad (4)$$

Where k represents the index of the cluster, Sk is the kth cluster set, lk denotes the centroid point in cluster Sk, which is also treated as the representative patients of the cluster, and K is the total number of the clusters.

There is a need to normalize the data point for eliminating the impact of the distinctive feature scales. To prepare the centroids used in building the cluster, the K-means method repeatedly adjusts the centroid location to decrease the Euclidean distance [18]. The number of clusters was determined using a similarity measurement [18]. The similarity metrics to assess the quality of clusters are shown in Eqs. (4 and 5) as follows:

$$d_{avg} = \frac{\sum_{k=1}^{K} \sum_{i \epsilon S_k} \sqrt{\sum_{j=1}^{F} (x_j^i - x_j^{\mu_k})^2}}{N} \quad (5)$$

$$d_{avg} = min \left[\frac{\sqrt{\sum_{j=1}^{F} \left(x_j^{\mu_{k_1}} - x_j^{\mu_k} \right)^2}}{N} \right], \forall k_1 \neq k_2 \quad (6)$$

Where davg denotes for average distance of each member i to the centroid lk in the same cluster Sk, Xi j denotes the jth input element of member I; dmin is the minimum distance between each two centroids, Xlk j represents the jth input element of centroid lk, N denotes the total number of data points, and F is the dimension of an input vector.

Noticeable, the K-means clustering method extracted five groups with a diverse number of samples and distribution of the extracted features using feature selection method. The best number of groups and clusters automatically fixed by the K-means algorithm with the assistance of the criterion specified in criterion group of the clustering. Figure 5 defines the results and clusters distribution.

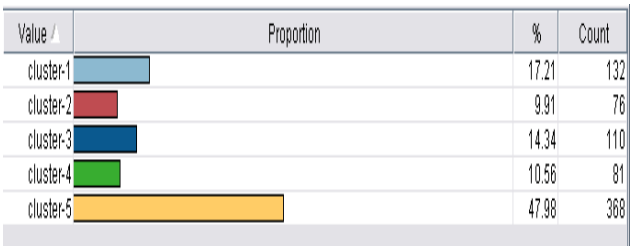


Fig. 5. Clusters distribution using K-means algorithm

Figure five above demonstrate the clustering distribution using k-means algorithm. The method generated five groups with different distribution samples and features. The distribution number of samples members are 767, 132, 76, 110, 81 and 368 for cluster1, cluster2, cluster3, cluster4, and cluster5, correspondingly. It's presented that the maximum numbers of samples scored by cluster 5 because of the similarity of the group instance features. Using these groups, the K-means clustering algorithm analyzed and defined the diabetes data; the clustering methods is used for describing the data. Thus, this research employed K-means technique to integrated with SVM prediction method. Based on the K-means clustering experimental on the diabetes dataset, the similar patients were grouped together and the different patients were clustered together too. The UCI-data clustering process is valuable at this stage to facilitate the diagnosis by classification and prediction in training and testing steps. In the classification phase, the SVM classifier is applied for obtaining a precise diagnosis of the patients. The general SVM method adopted for this problem to exploit the classifier as follows [23] and [24]:

$$maximize_x = \left[\sum_{i=1}^{n} x_i - \frac{1}{2} \sum_{ij=1}^{n} x_i x_j y_i y_j k(x_i, x_j) \right] \quad (7)$$

$$subject\ to \sum_{i=1}^{n} x_i y_j = 0, \ 0 \leq \forall x_i \leq L. \quad (8)$$

Where x stands for the learning vector, y denote the relationship label between the learning vectors; a denotes the variables vector of hyperplane classifier; K is a kernel function for measuring the distance between the learning vector xi and xj, and L stands for a drawback parameter to manage some misclassifications.

For example, if L is infinity, the predictor supplies an infinite penalty on misclassification to avoid misdiagnosis from taking place. A higher L ensures greater accuracy on learning data; simultaneously, it takes extra time to obtain the predictor. A lower L offers additional flexibility on the predictor on the tolerance of error. For SVM and K-mean clustering, a performance index employed to prove the accuracy of the introduced technique. The variables for a hybrid K-SVM method that is to be used in the experiments is measured as a constant change optimization procedure that is conceded out by the SVM method. Each partition used 50%, 60% and 70% as the training dataset and 50%, 40%, and 30% as the testing dataset using K-SVM as the prediction technique. The results of clustering process used as an input to the SVM classifier.

IV. Experimental Results and Discussion

The developed research used the mentioned experiment using SVM method before and after combination to examine the enhancement of the proposed approach. The obtained diagnosis results before improvement using SVM only are 77.09%, 75.27%, and 77.32% for learning data and 81.01%, 80.39%, and 77.29% for testing data respectively. On the other hand, the diagnosis accuracy results after improvement by using the hybrid method between K-means and SVM algorithm are 99.74, 99.78, and 99.81 for training data and 99.82, 99.85, and 99.90 for testing data respectively too. The evaluation results have been calculated as:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FP}) + (\text{TP} + \text{FN})} \times 100 \quad (9)$$

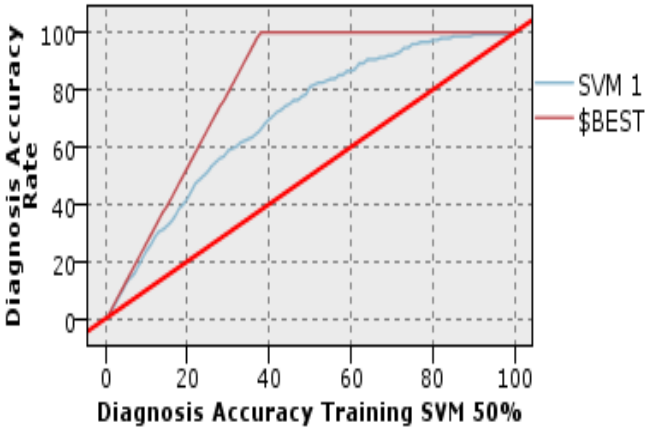
Where, True Positive (TP): The number of diabetics and non-diabetic patient's executable correctly classified. False Positive (FP): The number of diabetic executables classified as non-diabetic. True Negative (TN): The number of diabetics and non-diabetic patient's executable incorrectly classified. False Negative (FN): The number of non-diabetic executables classified as diabetic.

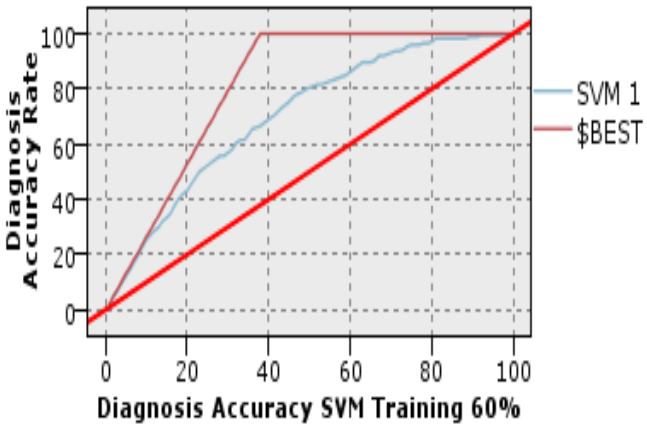
A. First Experiment

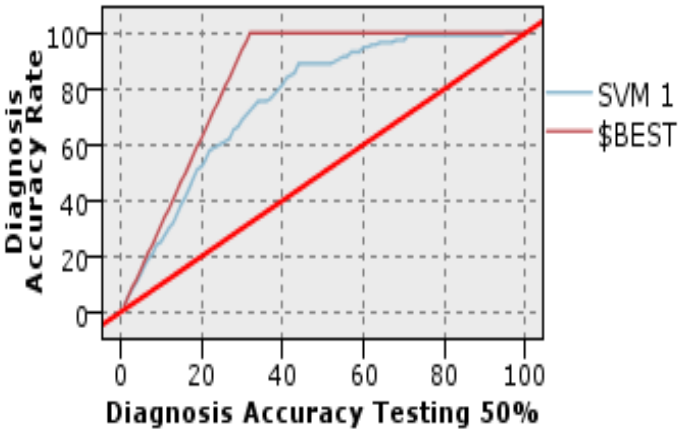
In this phase, the quantity of detecting diabetes patients (diabetic or non-diabetic) from the original UCI dataset was investigated. These experiments conducted on 768 patients reported and used by [7,12-14], had each sample described as either a diabetic or non-diabetic case. The experiments applied across 70, 60, and 50 instances as training experiments. Each learning and testing phase chose original diabetes patients variable as input features in SVM. Then the class variable is a target element (diabetic or non-diabetic). Additionally, it by using K-SVM, and diagnosis accuracy enhanced rather than used the SVM alone. The obtained results using SVM in learning and testing phases demonstrated in Table3. While the improved results based on SVM-K-means and important features presented in Table 4.

TABLE II. Results on the Original Pima Indian Diabetes Dataset using SVM Algorithm

Algorithm	Accuracy		Error	
	Training	***Testing***	***Training***	***Testing***
SVM	77.09 (50% size)	81.019(50% size)	22.91	18.99
SVM	75.27 (60% size)	80.39 (40% size)	24.73	19.61
SVM	77.32 (70% size)	77.29 (30% size)	22.68	22.71







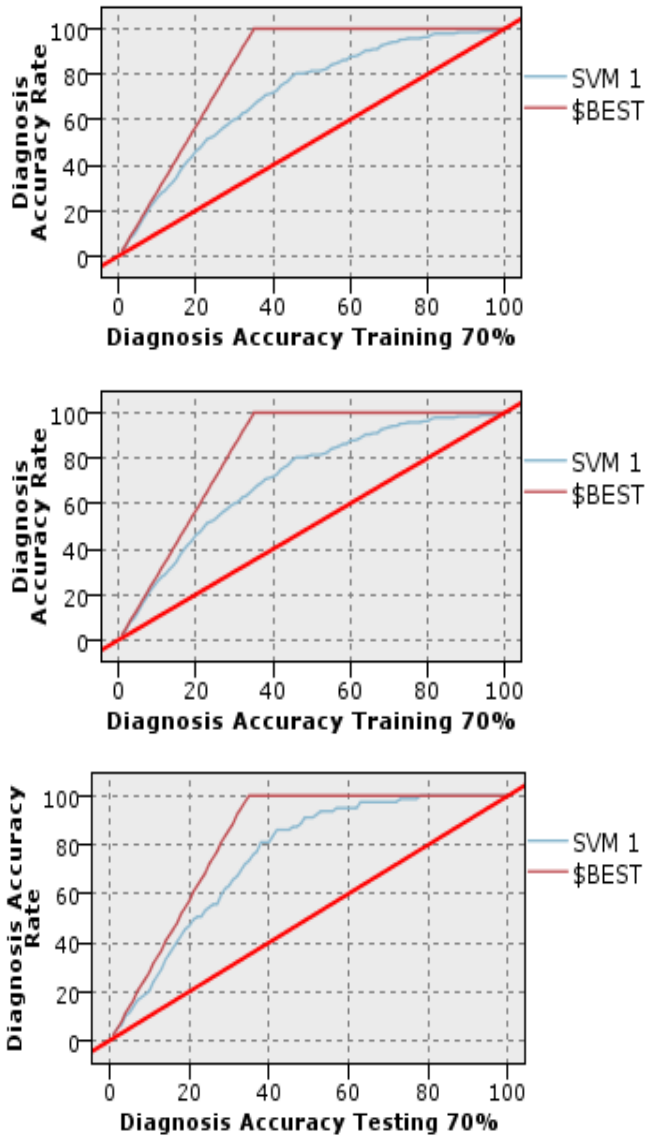


Fig. 6. Diagnosis Accuracy of SVM 50%, 60% and 70%

Figures 6 demonstrate the accuracy of the results of SVM algorithm only along with the best class in training and testing experiments. This is the gained charts for both training and testing output of the SVM before using the K-mean algorithm. The Classification using SVM for training data obtained a good diagnosis (77.32%) with 70% data size and 81.01% diagnosis accuracy for testing data with 50% data size. Gains chart with baseline, best line ($BEST-class) as well as the result of SVM before improvement is ($S-class). The accuracy ratio is presented in the X-axis. Using the SVM classifier, if the accuracy results are greater than the baseline with red color, the results are accepted; otherwise, the results are classified as rejected. On the other hand, the Y-axis represents the nominated prediction diagnosis between all the input variables for diabetes patinas. Expressively, diagnosis accuracy results that were achieved by the SVM are 77.09%, 75.27%, and 77.32% for learning experiments with amount 50%, 60%, and 70%, respectively, and 81.01%, 80.39%, and 77.29%for testing experiments with amount 50%, 60% and 70%, respectively.

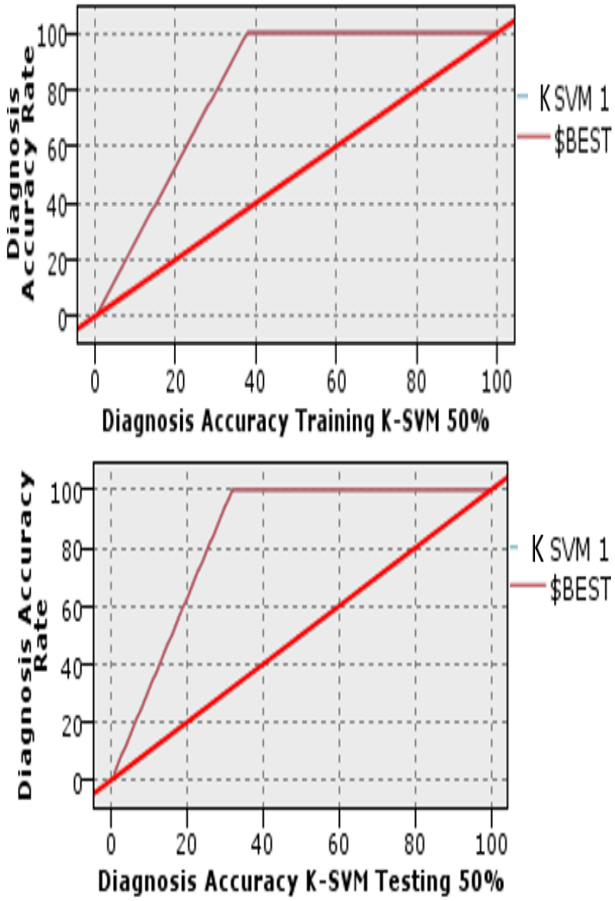
B. Second Experiment

The achieved accuracy after integration SVM and K-means is described in Figure 2 and Table4. The accuracy obtained during the diagnosis procedure is clarified by the results values.

As shown in Table 3, there are different results that were extracted with SVM. These results enhanced using K-SVM approach. Compared with Table3, the K-SVM approach obtained better accuracy than the SVM.

TABLE III. RESULTS ON THE ORIGINAL PIMA INDIAN DIABETES DATASET USING K-SVM ALGORITHM

Algorithm	Accuracy		Error	
	Training	*Testing*	*Training*	*Testing*
K-SVM	99.74(50% size)	99.82(50% size)	0.26	0.18
K-SVM	99.78(50% size)	99.85(50% size)	0.22	0.15
K-SVM	99.81(50% size)	99.90(50% size)	0.19	0.10



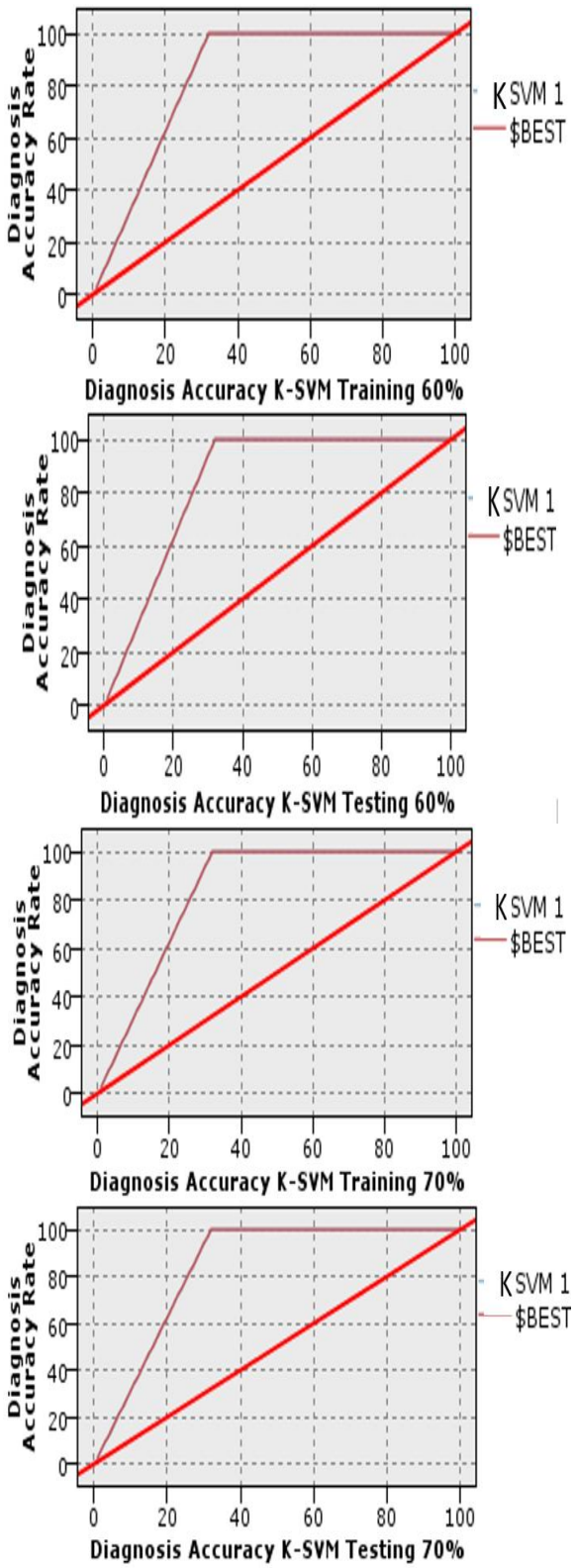


Fig. 7. Diagnosis Accuracy of SVM 50%, 60% and 70%

Figures 7 demonstrate the accuracy of the results of K-means-SVM method based on important features with the best class in training and testing experiments. The gained charts represented both training and testing results of the SVM with original dataset after using the K-mean algorithm with important features. The Classification using K-SVM for training data obtained optimal diagnosis (99.81%) with 70% data size and (99.90%) diagnosis accuracy for testing data with 70% data size. Gains chart with baseline, best line ($BEST-class) and the result of K-means-SVM after improvement is ($S-class). The X-axis shows the accuracy ratio. Based on the K- SVM technique, if the results are greater that the baseline with red color, then these results is accepted, otherwise the results will be categorized as unaccepted. On the other hand, the Y-axis represented the selected diagnosis accuracy between the selected features from diabetes patinas, containing Cluster type as a new feature produced from the clustering phase. It should be observed that the achieved diagnosis results using K-SVM are 99.74, 99.78, and 99.81 for learning experiments with amount 50%, 60%, and 70%, respectively, and 99.82, 99.85, and 99.90 for testing experiments with amount 50%, 60%, and 70%, respectively. It is also concluded that these results are improved than the SVM results.

Many researchers used the T-test statistical significance technique for methods comparison[27]. The t-tests statistical significance test performed between the results obtained from experiment 1 using SVM and experiment 2 using the hybrid method, and they showed improvements achieved by the K-SVM technique. Table 5 shows the standard deviations, some cases, standard errors, mean values, and significance results for the pairs of variables before and after improvement with K-means-SVM method (SVM, K-means-SVM) compared with the Paired Samples T-Test procedure. The Paired-Samples T-Test procedure compares the means of two variables that represent the same group at different times. The mean values of the two variables (SVM, K-means-SVM) are displayed in the Paired Samples Statistics Table. Since the Paired Samples T-Test makes a comparison of the means of the two variables, it is important to know what the mean values are. A little significance value for the T-test (typically less than 0.05) indicates that there is a difference between the two variables. The obtained t-test result is (0.003), this situation was emphasized in estimation measures, which means the K-SVM technique achieved significant results on the test accuracy.

TABLE IV. T-TEST COMPARISON RESULTS

	Variances result in 70%, 60%, and 50% dataset between the SVM and K-SVM					t	df	Sig. Value
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
SVM& K-SVM	20.29	2.03245	1.17343	15.2444	25.3422	17.2	2	.003

Different proposed diabetes diagnosis methods such as Decision tree[28], Discriminant Fuctions[29] and [30], and Bayesian network[31] examined and compared with the developed technique, and other diagnosis methods demonstrate in Figure 8.

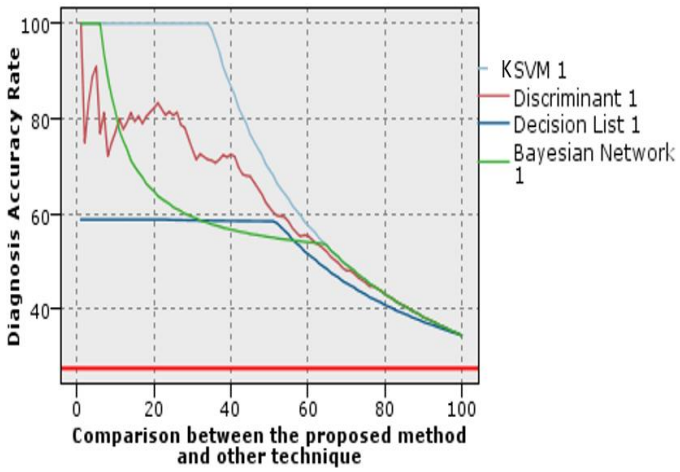


Fig. 8. Comparison between KSVM and other methods

In this paper, an integrated between K-means clustering algorithm and SVM method was proposed and discussed. The proposed technique analyzed and mined diabetes dataset for diagnosis purpose. Feature selection algorithm offered significant advantages when it came to generating important features probably[32]. The used to extract the essential features in many biomedical types of research such as cancer[33] and hepatitis[34]. The proposed method used the K-means clustering algorithm to improve the SVM method. Only the most important features as selected by feature selection method were utilized in the diagnosis and classification process. The results from the experimental tests against the UCI Pima Indian dataset showed that the overall of the proposed method performance achieved better results. The hypothesis presented the idea that the quality of diagnosis can be improved using K-SVM technique. The focus of the proposed method was adjusted so that only the most important features received attention. The T-Test was performed to examine the improvements achieved by the proposed method before and after combination process between K-means and SVM. The results of the T-Tests discovered the benefits of the proposed method discussed in this paper were statistically significant.

V. Conclusion and Future Work

This study tried to solve the problem of incorrect diagnosis problem of diabetes disease. The research proposed an integration technique between the SVM and K-means clustering mechanism for predicting an accurate diagnosis of diabetes disease. The study examined the combined method using the UCI Pima Indian diabetes standard dataset. A scientific experiment carried out to investigate the diagnosis accuracy for possible enhancement using The hybrid K-SVM. It has been proving that the integration between the K-means clustering method and SVM can enhance and produce accurate diagnosis results in diabetes disease. In the future work, the authors will integrate one of the optomization techniques for potintial enhancment.

Acknowledgment

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (830/125/D1437). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

References

[1] Abegunde, Dele O., Colin D. Mathers, Taghreed Adam, Monica Ortegon, and Kathleen Strong. "The burden and costs of chronic diseases in low-income and middle-income countries." The Lancet 370, no. 9603 (2007): 1929-1938.

[2] James, J.T., A new, evidence-based estimate of patient harms associated with hospital care. Journal of patient safety, 2013. 9(3): p. 122-128.

[3] Nanda, Surabhi, Mina Savvidou, Argyro Syngelaki, Ranjit Akolekar, and Kypros H. Nicolaides. "Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks." Prenatal diagnosis 31, no. 2 (2011): 135-141.

[4] Alssema, M., D. Vistisen, M. W. Heymans, G. Nijpels, Charlotte Glümer, P. Z. Zimmet, J. E. Shaw et al. "The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes." Diabetologia 54, no. 5 (2011): 1004-1012.

[5] Bennetts, Craig J., Tammy M. Owings, Ahmet Erdemir, Georgeanne Botek, and Peter R. Cavanagh. "Clustering and classification of regional peak plantar pressures of diabetic feet." Journal of biomechanics 46, no. 1 (2013): 19-25.

[6] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5, no. 5 (2013): 241-266.

[7] Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences 25, no. 2 (2013): 127-136.

[8] Kalaiselvi, C., and G. M. Nasira. "Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm." International Journal of Computing Algorithm (IJCOA) 4 (2015): 1403-1407.

[9] Saudek, Christopher D., William H. Herman, David B. Sacks, Richard M. Bergenstal, David Edelman, and Mayer B. Davidson. "A new look at screening and diagnosing diabetes mellitus." The Journal of Clinical Endocrinology & Metabolism 93, no. 7 (2008): 2447-2453.

[10] Lee, Tian-Shyug, Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines." Computational Statistics & Data Analysis 50, no. 4 (2006): 1113-1130.

[11] Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse." Artificial intelligence in medicine 26, no. 1 (2002): 37-54.

[12] Giveki, Davar, Hamid Salimi, GholamReza Bahmanyar, and Younes Khademian. "Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search." arXiv preprint arXiv:1201.2173 (2012).

[13] Acharya, U. Rajendra, Hamido Fujita, Shreya Bhat, Joel Ew Koh, Muhammad Adam, Dhanjoo N. Ghista, Vidya K. Sudarshan et al. "Automated diagnosis of diabetes using entropies and diabetic index." Journal of Mechanics in Medicine and Biology 16, no. 01 (2016): 1640008.

[14] Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." IEEE transactions on information technology in biomedicine 14, no. 4 (2010): 1114-1120.

[15] Yuwono, Mitchell, Ying Guo, Josh Wall, Jiaming Li, Sam West, Glenn Platt, and Steven W. Su. "Unsupervised feature selection using swarm intelligence and consensus clustering for automatic fault detection and diagnosis in heating ventilation and air conditioning systems." Applied Soft Computing 34 (2015): 402-425.

[16] Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. "Inductive learning algorithms and representations for text categorization." In Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155. ACM, 1998.

[17] Fagard, Robert H. "Exercise characteristics and the blood pressure response to dynamic physical training." Medicine and science in sports and exercise 33, no. 6; SUPP (2001): S484-S492.

[18] Jin, Jian-Ming. Theory and computation of electromagnetic fields. John Wiley & Sons, 2011.

[19] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31, no. 8 (2010): 651-666.

[20] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients." International Journal of Engineering and Advanced Technology 1, no. 3 (2012): 147-151.

[21] Al-Harbi, Sami H., and Victor J. Rayward-Smith. "Adapting k-means for supervised clustering." Applied Intelligence 24, no. 3 (2006): 219-226.

[22] Ordonez, Carlos. "Clustering binary data streams with K-means." In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pp. 12-19. ACM, 2003.

[23] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert Systems with Applications 40, no. 1 (2013): 200-210.

[24] Xing, Eric P., Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. "Distance metric learning with application to clustering with side-information." In NIPS, vol. 15, no. 505-512, p. 12. 2002.

[25] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.

[26] Danese, Daniele, Salvatore Sciacchitano, Antonella Farsetti, Mario Andreoli, and Alfredo Pontecorvi. "Diagnostic accuracy of conventional versus sonography-guided fine-needle aspiration biopsy of thyroid nodules." Thyroid 8, no. 1 (1998): 15-21.

[27] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12, no. 5 (2012): 1493-1502.

[28] Kaur, G. and A. Chhabra, Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 2014. 98(22).

[29] Polat, K., S. Güneş, and A. Arslan, A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 2008. 34(1): p. 482-487.

[30] Chien, B.-C., J.-Y. Lin, and W.-P. Yang, A classification tree based on discriminant functions. Journal of information science and engineering, 2006. 22(3): p. 573-594.

[31] Kumari, M., R. Vohra, and A. Arora, Prediction of Diabetes Using Bayesian Network. 2014.

[32] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications, 41(4):1476-1482, 2014.

[33] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2):3240-3247, 2009.

[34] L. F. Chen, C. T. Su, K. H. Chen, and P. C. Wang, "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis," Neural Computing and Applications, vol. 21, no. 8, pp. 2087–2096, 2012.

[35] Vasilakos, Athanasios V., Yu Tang, and Yuanzhe Yao. "Neural networks for computer-aided diagnosis in medicine: A review." Neurocomputing 216 (2016): 700-708.

[36] K.V. Dalakleidi, K. Zarkogianni, V.G. Karamanos, A.C. Thanopoulou, K.S.A.Nikita, A hybrid genetic algorithm for the selection of the critical features forrisk prediction of cardiovascular complications in Type 2 Diabetes patients, IEEE 13th International Conference on Bioinformatics and Bioengineering(2013) 1–4.

[37] Zheng, Tao, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International Journal of Medical Informatics 97 (2017): 120-127.