An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique

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

***Abstract*—This paper proposes an automatic diagnostic method for breast tumour disease using hybrid Support Vector Machine (SVM) and the Two-Step Clustering Technique. The hybrid technique is aimed at improving the diagnostic accuracy and reducing diagnostic miss-classification, thereby solving the classification problems related to Breast Tumour. To distinguish the hidden patterns of the malignant and benign tumours, the Two-Step algorithm and SVM have been combined and employed to differentiate the incoming tumours. The developed hybrid method enhances the accuracy by 99.1% when examined on the UCI-WBC data set. Moreover, in terms of evaluation measures, it has been shown experimentally results that the hybrid method outperforms the modern classification techniques for breast cancer diagnosis.**

Keywords*—Two-Step Clustering; Breast Cancer; SVM classification; Diagnosis; Tumors

I. Introduction

Now-a-days, Breast cancer is one of the serious dilemma facing the radiology scientists. Indisputable information is not available, but rather it was assessed that the newly malignancy examples in 2012 is more than 1,600,000 whereas the number of tumour passing away would spread more than 570,000 [1]. Breast cancer represented 29% of assessed new womanly tumour patients (790,740 patients), making it the most regularly diagnosed malignancy among ladies[1]. Diagnosis of cancerous cells in the breast is one of the biggest real-world medical problems. The diagnosis has always been a major problem in the medical field, based on various tests conducted on various patients. Tests are meant to aid the physician in making a proper and accurate diagnosis. However, miss-diagnosis sometimes occurs, especially in tumour and cancerous cells since it can be difficult to make an accurate diagnosis, even for a medicinal cancer expert. One of the drifting issues in the medicinal field is a diagnosis of the tumours. Mass descriptive tumour information and feature data on cancer studies can now be obtained with the aid of information technology. Mammography by radiologists and physicians has long been the means of predicting breast cancer. In 1994, ten radiologists analyzed and interpreted 150 mammograms to classify the tumour categories in the breasts [2]. The variation of the radiologists' clarifications brought on a low accuracy of diagnosis even though the value of using mammograms was proven. Above 89.5% of radiology scientists identified less than 3% of tumors from the study.

The remaining of the manuscript is sorted out as pursues. Section2 discusses the related literature review. In Section 3 and Section 4, the concept of the SVM and Two-Step Clustering algorithm. Section 5 provides a description of the involved idea of a hybrid technique. Section6 discusses Dataset. Experimental design of the hybrid approach was described in Section7. Section8 deals with results discussion of the introduced approach. Finally, section9 discusses the Conclusions and future work.

II. Related Works

Several technologies are now employed for the collection and analysis of the datasets. Given the large volume of cancer cases available, it is difficult for a medical doctor to absorb every particular feature of cancer. Accordingly, physicians increasingly depend on data analysis methodologies for making decisions related to cancer diagnosis. Some researchers are also relying on machine learning methods and data mining techniques (DM) for predicting and classifying breast tumour to accuracy improvement and deal with the increasing tumour information and feature data. A broad mixture tool and software for knowledge discovery (KD) behind large-scale data, DM is highly appropriate in the real world. Machine learning approaches and data mining were employed into a software-assisted system for breast tumour diagnosis by Wolberg [3], and Pena-Reyes [4]. Knowledge discovery techniques were used in tumour classification with positive results as demonstrated by the results of Pena-Reyes and Sipper's research, and the current breast tumour diagnosis became an obstacle of classification in the knowledge discovery domain. The current tumour feature datasets were classified separately into benign and malignant groups. Evaluating the classifier made it possible for a new incoming tumour to be classified, using the tumour's historical data, by finding out a classifier to divide the tumours types. Based on the related work a data mining methods were used to diagnose tumour based on cancer features data. The computational time increases as the number of descriptive tumour features increases. Different approaches in recognizing tumour shapes and getting the needed information and data for breast tumour diagnosis are investigated in this study to work with a huge amount of tumour attributes. This research conducted to find an accurate and efficient approach to analyzing and detect the incoming tumour pattern with the assist of knowledge discovery methods. Because the tumour features can be described in great detail, the unnecessary details lead to a

higher computational time for complex estimation without much influence to the concluding predictor. It is significant to note that a part of the basic requirement for cancer diagnosis also includes time complexity. In addition to time efficiency, a way of mining and extracting the essential data from the tremendous information, filtering the features and predicting the classification of the new tumour instances accurately is now of concern. Sequential backward and forward search to choose the most important mixture of features using the multilayer perceptron neural network to classify tumours had been proposed by Nezafat et.al [5]. F-score for finding the DNA virus discrimination was proposed for the selection of the best subset of DNA diseases for breast tumour analysis based on SVM [6] and [7]. Combining an SVM-based approach with feature reduction technique for diagnosis of breast tumour was proposed by Akay [8]. Through the use of the F-score for the measuring of the feature discrimination, the study conducted a consuming time for the optimal parameters adjusting mixed on the precision diagnosis to nominate the best subclass of the basic tumour attributes for learning stage by support vector machine[7]. Prasad, Biswas, and Jain [9] proposed a another combination method of SVM and heuristics to find out the significant attributes subclass for SVM learning stage rather than the extensive search. In addition to an improvement in diagnosis cancer accuracy, their results also combined with PSO technique to cut down the time complexity for the significant training due to the inference on dimension feature space and searching for the best feature. Accurate diagnostic measurements are taken on the FNA [10] in this dataset. These statistics claim that compared to other forms of cancer, breast tumour places the 3rd position among diagnosed new patients, the first and second places being occupied by genital organs and digestive systems tumour, respectively. The surgical biopsy is the best approach for confirming malignancy with high-level sensitivity in the prognosis of the disease and breast tumour diagnosis. But it is an expensive operation with a negative impact on the patient's psychology. Dubey, A. K., U. Gupta, et al. [20] investigated the influences of k-means clustering method based on distance, centroid, epoch, split method, and iteration to identify and consider the integration of computational measures for possible extract highly accuracy of breast cancer Wisconsin dataset. Their method obtained 92% accuracy in term of precision. Zheng, B., S. W. Yoon, et al.[21] proposed a breast cancer diagnosis method using a combination between K-means and SVM technique. The method used feature selection technique to extract the important feature for the potential to improve the diagnosis accuracy. The hybrid approach achieved 97.38% accuracy results in term of evaluation measures. S. Aruna et al. [22] examined different classification methods such as SVM-RBF kernel, RBF neural networks, simple CART and J48 methods to extract the optimum classifier in WBC dataset. The experimental output proves that SVM-RBF kernel is better than other classification techniques with 96.84% accuracy scores in WBC dataset.

Considering the aforementioned research work, there is a shift towards less invasive data mining methods which can give the same rates of accuracy without the problems associated with surgical biopsy. The current study proposed a hybrid SVM-Two-Step method as one of the diagnostic solutions when the aim is to classify and predict the patient's medical information to detect the level of breast cancer accurately. Identifying a difference between malignant and benign cancer is the function of diagnosis. Once the cancer is diagnosed, it is necessary to follow with a classification of the expected the disease course. More importantly, the Two-Step Clustering methodologies and an efficient probabilistic support vector machine are also studied in this research.

The different between the TwoStep-SVM method and other classification techniques is that the hybrid approach used a clustering output as an input feature which it can help for improving the diagnosis accuracy. Where the other classifiers used an original dataset feature as an input. Another difference is that the twoStep clustering algorithm can group the breast cancer dataset automatically to select the similar samples and features. Thus it can increase the correlation between the input features to assist a classifier in extracting accurate results.

III. SUPPORT VECTOR MACHINE (SVM)

The SVM is a modern technique rapidly gaining popularity as a result of the helpful results that have been accomplished in a widespread diversity of data mining issues, due to their strong hypothetical, theoretical underpinnings in statistical learning theory [11], [12] and [13]. SVM is a binary classification method according to the theory of statistical learning that has been used with much success in different defy on large datasets and nonlinear classification problems [14] and [15]. It has been found valuable in solving linear separated (LS) as well as non-linear separated problems (NLS) [16]. SVM predictors use the hyper-plane in separate classes. Each hyper-plane is defined by its direction (w), the exact location in space or a threshold is (b), (xi) is the input vector of element N or text content and indicates the class. A group of the training cases is presented by equations 1 and 2.

$$(x1, y1), (x2, y2), \ldots, (xk, yk);\ Xi \in R^d \qquad (1)$$

k is the training dataset number and d represents the number of dimensions of input dataset: $yi \in \{-1,+1\}$; $i = 1, 2, \ldots, k$. The decision function of the form Eq. 2.

$$f(x, w, b) = \operatorname{sgn}((w \cdot xi) + b), w \in R^d, \quad b \in R \qquad (2)$$

The margins are the region among the hyper-plane, which separates two classes; the margins demonstrate the classification of breast cancer by SVM. Let the distance from the adjacent data point to the hyper-plane be $\frac{1}{||w||}$. There is one optimum separating hyper-plane among separate hyper-plane, and the distance of two SV points from diverse sides of this hyper-plane is maximal. Then the vertical distance from the origin to this hyper-plane is $\frac{1}{||w||}$, and the margin distance separating hyper-plane is $\frac{2}{||w||}$. The minimum distance of the margin is equivalent to $\frac{1}{2}||w||^2$ (named primal problem) and getting the maximum potential margin is the primary knowledge of SVM method. Figure 1 demonstrates the prediction of breast cancer utilizing Support Vector Machine.

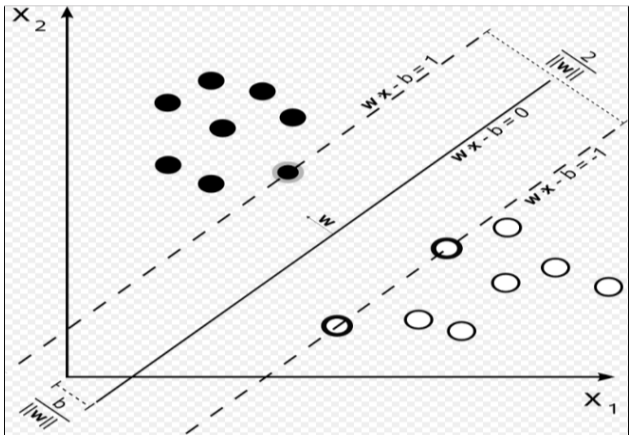

Fig. 1. Classification of breast cancer using SVM

IV. Two-Step Clustering Algorithm

Several approaches used Two-step clustering algorithm in different felids such as [23,24, and 25]. Najjar, A et al. [23] proposed an exploring analytics method extending from Smyth study [24] for analyzing healthcare data. The proposed method used the two-Step clustering algorithm for heterogeneous finite mixture model using two-step; firstly, including a joint mix of multinomial distribution and Gaussian to handle categorical and numerical inputs. Then, featuring a mix the hidden Markov model to handle orders of categorical input. The method is measured on a real-world system and the obtained good results for identifying health services with big families. V. Deneshkumar et al. [25] proposed a method for detecting the outlier and defining the effect influence in the diabetic people using two-step clustering algorithm and different data mining methods. The method tried to find the patterns and relationships within large medical data to extract new clinical knowledge.

The TwoStep Clustering algorithm is associated wildcat method proposed to disclose natural clusters (or groups) inside a knowledge set that might or not be obvious [17]. The method utilized by this process has many different options that distinguish it from ancient clustering methods:

- The capability of make groups of elements that can support each continuous and categorical variables.
- Determine the number of clusters automatically.
- Analysis a big corpus efficiently.

A. Principles of Clustering

To handle continuous and categorical variables, the TwoStep Clustering algorithm applies a likelihood distance measure that supposes that variables within the group model are freelance. Additionally, every categorical variable is intended to own a multinomial distribution, and every continuous variable is expected to own a standard (Gaussian) allocation. Experimental internal testing specifies that the process is properly strong to violations of each the belief of individuality and therefore the spatial arrangement assumptions. The two-steps of the algorithm's rule are concluded as:

- **phase 1.** The process starts with the development of a Cluster Features (CF) Tree. The tree starts by putting the first case at the tree root in a leaf node that conveys variable information for that case. Every consecutive case is then supplemental to associate existing node or forms a new node according to the distance similarity between the existing nodes.
- **phase 2.** By using agglomerative clustering (AC) mechanism, the leaf nodes of the Cluster Features tree are then clustered. The AC can be employed to range the produced solutions. The best number of clusters can be determined by comparing these clusters using the Akaike Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC).

To define the similarity score between the object, the Euclidean distance measure is used in a proposed hybrid method as show in Eq. 3.

$$\text{Dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

A Euclidean vector is the position of a point in a likelihood n-space. Therefore, *X* is (*Xn, Xn,. . . , Xn*) and Y is (*Y1, Y2,. . . , Yn*) are likelihood vectors, beginning from the origin of the space, and two points are indicated by their tips [18].

V. Hybrid Approach

The introduced technique is a hybrid method for breast cancer dataset prediction using Two-Step clustering and SVM methods and consists of two sub-methods: Two-Step data clustering based on features similarity using likelihood distance measure, and classification breast cancer dataset based on the SVM algorithm. The purpose of this research is to introduce a cancer diagnostic classification approach with the aid of a hybrid Two-Step data clustering algorithm and the SVM prediction method for the enhancement of the classification accuracy (effectiveness) and to reduce the rate of misclassification. This work pioneers a new approach which combines the supervised and the unsupervised learning methods Two-Step clustering algorithm and SVM techniques. A qualified research has been conducted on the SVM classification and Two-Step data clustering structure on breast cancer features. Then the results of clusters used as inputs to the prediction method using the SVM technique as classifiers for cancer cases. The Hybrid TwoStep-SVM technique is considered to investigate the result of the trained method. As a result of a large number of cases correlated with the cancer data, The dataset was split into ten parts as 10-folds cross-validations for training and testing the TwoStep-SVM method. **Figure 2** shows the stages of the introduced technique (TwoStep-SVM stages).

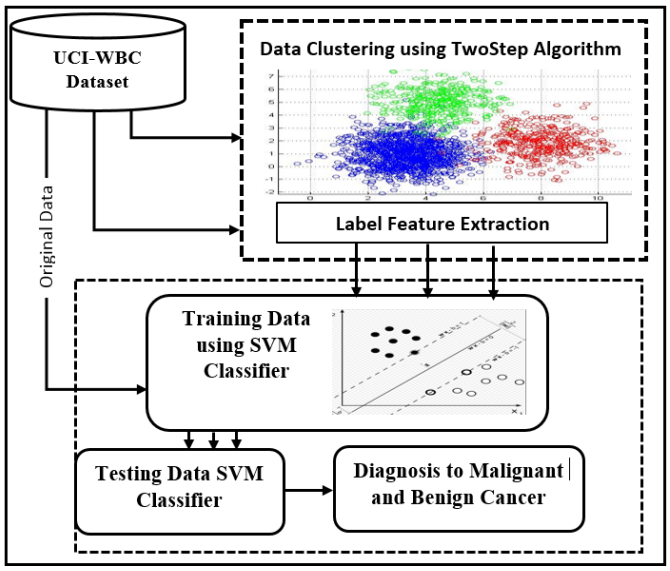

Fig. 2. Hybrid TwoStep-SVM Approach

VI. DATASET

This research was conducted based on the Wisconsin Breast Cancer (WBC) corpus. The corpus was widely used as [7,8,9,20,21 and, 22] and always used to discriminate cancerous (malignant) from the non-cancerous (benign) sample. Table 1 gives a description of the WBC dataset. The WBC dataset is made up of a group of features with some instances and classification patterns. The number of cases and samples of this dataset is 699, with 11 features classified into two classes.

TABLE. I. FEATURES OF BREAST CANCER

No	Feature and Range
1	Sample code number: id
2	Clump Thickness:1 - 10
3	Uniformity of Cell Size: 1 - 10
4	Uniformity of Cell Shape: 1 - 10
5	Marginal Adhesion: 1 - 10
6	Single Epithelial Cell Size: 1 - 10
7	Bare Nuclei: 1 - 10
8	Bland Chromatin: 1 - 10
9	Normal Nucleoli: 1 - 10
10	Mitoses: 1 - 10
11	Class: (2 for benign, 4 for malignant)

In the group thickness, cancerous cells are usually gathered in multilayer, while benign cells tend to be assembled in monolayers. Whereas in the homogeny of cell shape and size, the cancerous cells tend to differ. That is why these parameters are important in defining if the cells are malignant or not. Healthy cells a tendency to stick together in Marginal adhesion cases, where cancerous cells most often lose this capability. Thus, the damage of adhesion is a sign of cancer. In the case of the single epithelial cells size, the size is associated to the uniformity stated above. Significantly, enlarged epithelial cells may be malignant cells. Nuclei that is not surrounded by cytoplasm (the rest of the cells) is called the bare nuclei. Typically, they are seen in benign cancer. The Bland Chromatin refers to a uniform "texture" of the nucleus seen in benign cells. The chromatin is often more coarse in tumour cells. The normal nucleoli are minor structures realized in the nucleus. In healthy cell, the nucleolus is typically small extremely, if observable at all. The nucleoli be more distinguished in cancer cells; occasionally there are further of them. Lastly, Mitosis is nuclear division plus cytokines which extract two duplicate daughter cells through prophase. The cell divides and replicates by this process. Counting the number of mitoses can enable pathologists to determine the grade of cancer.

VII. EXPERIMENTAL DESIGN

This research conducted an experimental design using UCI WBC breast tumour dataset for the assessment of the performance of introduced algorithm. To compare the accuracy of cancer predictors, experiments ran the *Two-Step*-SVM training method using 10-folds cross validation. The dataset was broken down into ten sets. Each set represented 10% from the original dataset to allow every slice of the dataset to take a turn as a testing data. For each round, the experiment used nine sets for training process and the reminder one for the testing process. The *Two-Step* technique is employed for grouping tumours based on similar class benign and malignant tumour features. From the explanation of the Two-Step clustering algorithm, the result of Two-Step algorithm extract 5 clusters with a different number of instances and features distributed from feature 1 to feature 11. The algorithm automatically determines the optimal number of groups with the assistance of the criterion defined in criterion cluster of the grouping. Table 1 describes the outcomes clusters while Table 2 illustrates the distribution of these clusters.

TABLE. II. RESULT OF INSTANCES DISTRIBUTION CLUSTERS BASED ON ALL BREAST CANCER FEATURES

Feature Name	Number of Instances for Each Cluster				
	1	2	3	4	5
ID	387	0	9	2	4
Normal Nucleoli	1	85	0	46	0
Clump Thickness	21	0	4	2	3
Uniformity of Cell Size	13	0	6	0	9
Uniformity of Cell Shape	5	0	10	0	4
Marginal Adhesion	8	0	10	2	10
Single Epithelial Cell Size	0	0	3	0	1
Bare Nuclei	0	0	2	0	6
Bland Chromatin	0	0	14	0	7
Mitoses	0	0	3	3	3
Class	12	0	3	0	1

In Table 2, TwoStep clustering algorithm results extracted 5 clusters or groups; these are all important clusters. The distributed numbers of instances members are 447, 55, 85, 64,

and 48-form cluster 1 to cluster 5 sequentially. It is shown that the highest number of instances due to the similarity of the member features is scored by cluster 1. The majority of the members of cluster 1 are similar in Bare Nuclei feature; Table 2 demonstrates the shared members of the other features in the cluster 1. The high score number of participated members among Clusters 3 and 5 is Mitoses feature with 85 and 46, respectively. In cluster 4 and 5, the Bland Chromatin feature scored with ten members and ranked as a high score among these cluster members. On the other hand, there is a less number of instances ranking to cluster 5 with 48 instances due to the variation and discrimination of cluster member features. Via these clusters, the Two-Step Clustering algorithm analyzed and described the breast cancer dataset; the main task of different clustering techniques is data description. The clustering algorithm was selected to be hybridized with SVM to enhance the classification and prediction process. The steps of how the clustering was used and combined with the SVM classifier are; first, the TwoStep method conducted to cluster the corpus of data into different groups. The output of these groups and clustering is represented in a new variable feature named label. The values of label feature are the cluster name such as cluster1, cluster2, etc. Each record in the dataset was labeled with the cluster name. Then, the SVM classifier was applied with the label feature for potential generating accurate diagnosis result with high prediction accuracy.

In the prediction part, the SVM algorithm is utilized for achieving accurate prediction cause of its high level of accuracy. Commonly, SVM algorithm adopted to find the predictor as follows [34] and [15]:

$$maximize_x = \left[\sum_{i=1}^{n} x_i - \frac{1}{2} \sum_{ij=1}^{n} x_i x_j y_i y_j k(x_i, x_j) \right] \quad (4)$$

$$subject\ to \sum_{i=1}^{n} x_i y_j = 0,\ \ 0 \le \forall x_i \le L. \quad (5)$$

Where x stands for the training vector, y represents the associated label between the training vectors, a denotes the variables vector of hyperplane classifier, K is a kernel function for assessing the distance between the learning vector xi and xj, and L stands for a penalty parameter to manage some misclassifications. For instance, if L is infinity, the predictor supplies an infinite penalty on classification error to prevent classification error from taking place. A higher L ensures a higher precision on learning dataset; simultaneously, it takes extra time to achieve the predictor. A lower L provides additional flexibility on the predictor on the tolerance of fault. In this situation, the results are not much affected by the different kernel function, and sigmoid kernel function has been utilized in the SVM method.

VIII. Experimental Result and Discussion

For the purpose of carrying out an experimental study, breast cancer dataset was learned. As stated earlier discussion, the research used 10-folds cross-validations approach for training and testing the dataset. The experiment applied cross the dataset using SVM classifier without clustering results and with clustering results to investigate the improvement outcomes of the hybrid approach. The cross-validations process obtained computed a diagnosis accuracy results as:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FP}) + (\text{TP} + \text{FN})} \quad (6)$$

Where,

True Positive (TP): The number of benign and malignant executables correctly classified; False Positive (FP): The number of benign executables classified as malignant; True Negative (TN): The number of benign and malignant executables incorrectly classified; False Negative (FN): The number of malignant executables classified as benign.

In the experiments, the WBC dataset was used in order to determine the breast cancer stage (benign or malignant). The dataset had each instance reported as either a benign or a malignant case. The hybrid technique applied by training and testing the dataset using hybrid TwoStep and SVM method. Using Two-Step algorithm, the dataset then was divided into different clusters with each cluster having different instances. The main objectives of clustering in this study is to extract patterns and structures by collecting the breast cancer samples with similar patterns together thus, the complexity will be reduced and the diagnosis interpretation will be accurate. The achieved performances from the training and testing process on the dataset are illustrated in Table 3 which demonstrates a set of results obtained by SVM classifier technique without clustering and with clustering using the TwoStep algorithm. In the combination process, the output of the TwoSetp is added as a new feature to label each instance in the dataset with a cluster name as discussed in Section 7. This feature can increase the correlation between the instances by grouping the dataset into different clusters, each with similar instances. The SVM classifier employed again with the output of the TwoStep method for possible obtaining high accuracy. A 10-folds cross-validations were applied in the trining and testing process with and without clustering. Each training and testing experiment uses breast cancer dataset features as an input variable to the SVM. Then the target field is a class feature (benign or malignant). The results of the SVM classifier with clustering proved an improvement performance when the SVM technique classified the dataset with TwoStep cluster output. Interestingly, the TwoStep clustering algorithm increases the diagnosis accuracy with 99.1% ratio as shown in Table 3.

Figures 3 and 4 demonstrate both training and testing output of the SVM without clustering and with clustering using Two-Step Clustering. The 10-folds cross-validations were calculated, and the average classification results using SVM without clustering obtained 96.19% and 95.23% for training and testing experiments respectively. The figures also represent the achieved results of the SVM classifier with clustering using a TwoStep algorithm with 99.39% in the training and 99.1 in the testing experiments. The training and testing high-performance results without clustering achieved in folds number 6 and 7 with 97.30% accuracy ratio.

TABLE. III. RESULTS ON THE WBC DATASET USING SVM TECHNIQUE

Fold No	SVM Accuracy Results Without Clustering		SVM Accuracy Results With Clustering Two-step Algorithm	
	Training Results %	Testing Results%	Training Results %	Testing Results%
Fold 1	96.8%	96.6%	99.30%	98.7%
Fold 2	95.5%	95.7%	99.40%	98.9%
Fold 3	94.9%	94.3%	98.90%	98.6%
Fold 4	95.3%	95.6%	99.60%	99.5%
Fold 5	96.6%	96.3%	99.80%	99.3%
Fold 6	97.3%	93.9%	99.30%	99.2%
Fold 7	96.4%	97.3%	99.50%	99.3%
Fold 8	96.3%	95.5%	99.20%	99.0%
Fold 9	95.9%	94.1%	99.30%	99.1%
Fold 0	96.9%	93.0%	99.60%	99.4%
Average	**96.19%**	**95.23%**	**99.39%**	**99.10%**

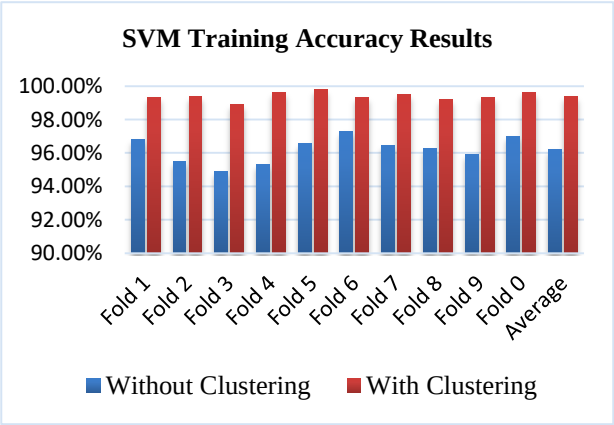

Fig. 3. SVM Training results

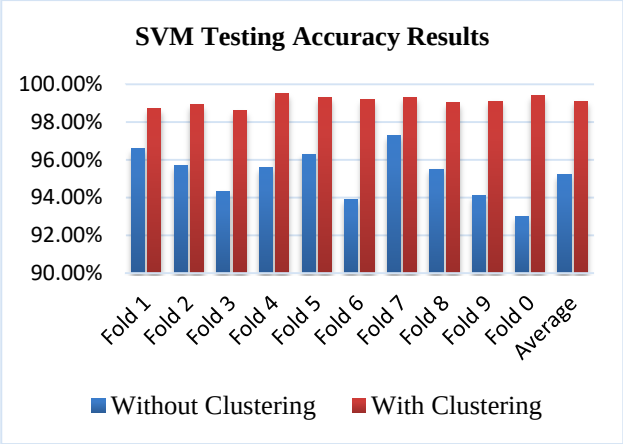

Fig. 4. SVM Testing results

On the other hand, the prediction results with clustering were achieved in folds number 5 and 4 in training and testing experiments respectively. The conclusion is that there is an enhancement while using the TwoStep clustering algorithm. The result of the SVM with clustering is better, and the breast cancer diagnosis is more accurate when using a combination of the SVM output with the Two-Step Clustering algorithm.

This research performed *T-test* algorithm as statistical significance test between the obtained results from the first experiment using SVM and second experiment using the TwoStep-SVM method, and it presented the enhancements obtained by using the Two-step-SVM technique. The small significance result for the T-test (normally below 0.05) specifies that there is a significant change between the two variables. Base on the obtained results in Table 4, regarding the accuracy of values (0.014), this condition was emphasized in assessment measures. It indicates that the TwoStep-SVM achieved significant improvement on the accuracy and there is a significant difference between the SVM with clustering and vice-versa.

The Comparison between Two-step-SVM technique and current approaches demonstrates in Figure 10. The proposed TwoStep-SVM obtained high accuracy value by 99.1.

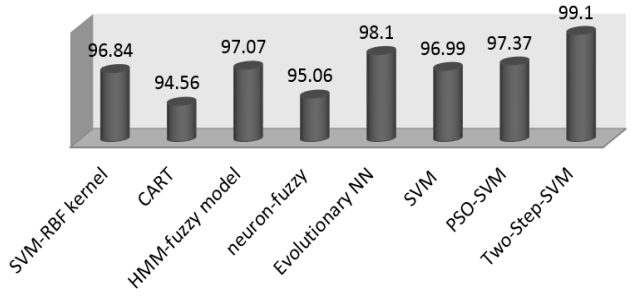

Fig. 5. Accuracy comparison between the Two-step-SVM technique and current breast cancer classifier method

TABLE. IV. T-TEST COMPARISON RESULTS

	Differences between accuracy result in 50%, 60%, and 70%, before and after the improvement					t	df	Significant Value
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
SVM & SVM-Two-Step	-0.0387	0.01409	0.000199	-0.048783	-0.028617	-8.68	18	0.000011

The time complexity of the hybrid TwoStep-SVM was also computed based on machine time efficiently. Due to the WBC dataset structure (Vector of data) that consist of number of columns (*n*) and number of rows (*m*), the time complexity can be computed as (*n*m*) and it is belong to the (*n*m*) class. Where (*n)* is represents the breast cancer features and (*m)* represents the breast cancer patients. The computational time of the previous studies such as [21] and [22] was calculated based on the machine time complexity (execution time/ CPU time) where the complexity time of the proposed hybrid method is calculated based on the mathematical computation.

IX. CONCLUSION AND FUTURE WORK

This research considered a major challenge in the health sector today the problems surrounding cancer disease detection. Significantly, this study has pioneered the use of the hybrid Two-Step-SVM method for predicting cancer disease. Moreover, the research investigated the hybrid algorithm on a WBC-UCI dataset which is a standard dataset utilized in the diagnosis of breast tumour. It has been proved that the SVM with the Two-Step algorithms can significantly improve the prediction accuracy rate and decrease the miss-classification error in cancer disease. More importanlly, the hybrid method improved the prediction accuracy following the methodology explained in section 6. In the future work, an optimization method will be combined with the SVM-two-step clustering algorithm to enhance the diagnosis accuracy.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (830/17/D1437). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

REFERENCES

[1] R. Siegel, et al., "Cancer Statistics, 2012," CA: a cancer journal for clinicians, vol. 62, pp. 10-29, 2012.

[2] J. G. Elmore, et al., "Variability in radiologists' interpretations of mammograms," New England Journal of Medicine, vol. 331, pp. 1493-1499, 1994.

[3] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative cytology and histology, 17(2), 77-87.

[4] Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. Artificial intelligence in medicine, 17(2), 131-155.

[5] Nezafat, R., Tabesh, A., Akhavan, S., Lucas, C., & Zia, M. (1998). Feature selection and classification for diagnosing breast cancer. In Proceedings of international association of science and technology for development international conference (pp. 310-313).

[6] Chen, Y. W., & Lin, C. J. (2006). Combining SVMs with various feature selection strategies. In Feature extraction (pp. 315-324). Springer Berlin Heidelberg.

[7] Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Systems with Applications, 34(1), 578-587.

[8] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

[9] Prasad, Y., Biswas, K. K., & Jain, C. K. (2010, June). SVM classifier based feature selection using GA, ACO and PSO for siRNA design. In International Conference in Swarm Intelligence (pp. 307-314). Springer Berlin Heidelberg.

[10] Cortes, J., O'Shaughnessy, J., Loesch, D., Blum, J. L., Vahdat, L. T., Petrakova, K., & Vladimirov, V. (2011). Eribulin monotherapy versus treatment of physician's choice in patients with metastatic breast cancer (EMBRACE): a phase 3 open-label randomised study. The Lancet, 377(9769), 914-923.

[11] Danese, D., Sciacchitano, S., Farsetti, A., Andreoli, M., & Pontecorvi, A. (1998). Diagnostic accuracy of conventional versus sonography-guided fine-needle aspiration biopsy of thyroid nodules. Thyroid, 8(1), 15-21.

[12] Hussain, H., Benkrid, K., & ŞEKER, H. (2016). Novel dynamic partial reconfiguration implementations of the support vector machine classifier on FPGA. Turkish Journal of Electrical Engineering & Computer Sciences, 24(5), 3371-3387.

[13] F. Salcedo-Campos, et al., "Segmental parameterisation and statistical modeling of e-mail headers for spam detection," Information Sciences, vol. 195, pp. 45-61, 2012.

[14] Casillas-Ramírez, A., Mosbah, I. B., Ramalho, F., Roselló-Catafau, J., & Peralta, C. (2006). Past and future approaches to ischemia-reperfusion lesion associated with liver transplantation. Life sciences, 79(20), 1881-1894.

[15] Zhou, C., Wu, Y. L., Chen, G., Feng, J., Liu, X. Q., Wang, C., ... & Lu, S. (2011). Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. The lancet oncology, 12(8), 735-742.

[16] Jordan, C. T., Guzman, M. L., & Noble, M. (2006). Cancer stem cells. New England Journal of Medicine, 355(12), 1253-1261.

[17] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

[18] Torres GJ, Basnet RB, Sung AH, Mukkamala S, Ribero BM (2009) A similarity measure for clustering and its applications. Int J Electr Comput Syst Eng 3(3):164–170.

[19] Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the national academy of sciences, 87(23), 9193-9196.

[20] Dubey, A. K., Gupta, U., & Jain, S. (2016). Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International journal of computer assisted radiology and surgery, 11(11), 2033-2047.

[21] Zheng, B., S. W. Yoon, et al. (2014). "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.

[22] Aruna, S., S. P. Rajagopalan, and L. V. Nandakishore (2011). "Knowledge based analysis of various statistical tools in detecting breast cancer." Computer Science & Information Technology 2: 37-45.

[23] Najjar, A., Gagné, C., & Reinharz, D. (2015, November). Two-Step Heterogeneous Finite Mixture Model Clustering for Mining Healthcare Databases. In Data Mining (ICDM), 2015 IEEE International Conference on (pp. 931-936). IEEE.

[24] P. Smyth. Probabilistic model-based clustering of multivariate and sequential data (1999). In Proceedings of the Seventh International Workshop on AI and Statistics, pages 299–304. San Francisco, CA: Morgan Kaufman.

[25] Deneshkumar, V., K. Senthamaraikannan, and M. Manikandan (2014), Identification of Outliers in Medical Diagnostic System Using Data Mining Techniques. International Journal of Statistics and Applications. 4(6): p. 241-248.