Received: October 21, 2020. Revised: December 14, 2020. 190

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer

Wathiq Laftah Al-Yaseen[1]* Ammar Jehad[1] Qusay Abdullah Abed[1]
Ali Kadhum Idrees[2]

[1]Kerbala Technical Institute, Al-Furat Al-Awsat Technical University, 56001, Kerbala, Iraq
[2]Department of Computer Science, University of Babylon, Babylon, Iraq
* Corresponding author's Email: wathiq@atu.edu.iq

Abstract: Breast cancer has been recently considered as one of the broadly spread diseases that causes death among women. Early disease diagnosis is a critical aim in building the treatment policies and is extremely related to safety of patient. Therefore, there is a necessity for computer aided detection (CAD) in order to provide accurate and rapid diagnosis for breast cancer. Recently, many classification models utilizing machine learning approaches have been adopted and modified to diagnose breast cancer disease. Moreover, the performance of each model depends on different compositions such as the number and type of data features and the parameters of model. In order to enhance the performance of classification model, this research proposes a model using modified K-means algorithm to create a new training dataset of breast cancer which can highly improve the performance of support vector machine model. A modified K-means algorithm is also proposed to build a high quality training dataset that contributes significantly to reduce the training time of classifiers, and improve the performance of classifier. The proposed model handles the noise and irregularity in data and produce high quality dataset which represents all the cases of disease. The two recognized datasets Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) have been used to examine and appraise the performance of the proposed model. The experimental results show that the proposed model has a significant performance compared to other previous works and with accuracy level of 98.067%, sensitivity of 100%, specificity of 94.811%, precision of 97.011% and finally with area under the curve related to the receiver operating characteristic of 97.406%.

Keywords: Breast cancer diagnosis, Machine learning, Support vector machine, K-means, Classification.

1. Introduction

Nowadays, the diagnosis of cancer is becoming one of the important issues that challenge researchers. Breast cancer is one of the most general cancers that is recognized to affect around 10% of women of the world at some periods of their life [1]. However, the researchers can detect the breast tumors at early stages based on combined of computer techniques (Computer Aided Diagnosis CAD) with biomedical technologies (such as X-ray radiography technology or B-ultrasonic) [2]. In the field of machine learning, the prediction of breast tumors is also considered as a classification problem that considers different breast tumor measurements instead of traditional diagnosis lab tests; for instance, positron emission tomography, breast biopsy and imaging of magnetic resonance. In the last years, various sophisticated methods based on machine learning techniques have been used to help the early diagnosis of breast cancer such as support vector machines (SVM) [3-5], multilayer perceptrons (MLPs) [6, 7] and logistic regression (LR) [8-10]. The stability and performance of these methods depend on many factors like the properties of features, the parameters of algorithms, and the structure of model. The challenges still exist to find the best way to attain the high performance. Moreover, the prediction of breast cancer is directly affected patients' treatment and safety, therefore, developing predicting models that are robust and reliable is so important and vital for those specialized in data

mining. SVMs exhibit good performance with breast cancer problem in terms of prediction of malignant and benign breast tumors.

K-means is among these algorithms which has been utilized in the diagnosis of breast cancer tumors. K-means is applied to cluster the dataset into similar groups in order to discover the meaningful patterns of gathering unlabeled data. Therefore, the samples in the same cluster have the same characteristics while those among clusters have different ones. Several research efforts such as [11-13] introduced K-means as an effective method for grouping the breast cancer dataset into clusters which involve benign and malignant cases. Furthermore, other studies such as [14, 15] have proposed combined methods, based on K-means and other techniques to build models for classify the breast cancer tumors. In 2018 [16], a comparative study of K-means and comparing it with fuzzy C-means algorithms is presented on the breast cancer. This study showed that K-means algorithm was more prominent and consistent in terms of computation time when comparing with FCM which required more time to implement several fuzzy calculations and iterations. Support Vector Machines (SVMs) have gained more attention through significant success in the last years. The SVM technique splits the collected data into several classes utilizing the hyperplane and by recognizing the highest margin among classes to ensure accurate results [17].

In this study, the researchers introduce a model based on modified K-means algorithm and SVM technique. Modified K-means algorithm has been employed to preprocessing the training dataset of breast cancer. Moreover, modified K-means reduces the number of training samples and produces a new training dataset which represent entirely original training dataset. The modification of K-means algorithm is by proposed a new manner for selecting the initial centroids of clusters which represent all cases. Moreover, the algorithm uses a method for selecting the centroids of clusters depending on a distance threshold. The distance threshold represents the maximum distance between centroids of clusters and the samples of dataset. For example, if the distance between a sample and the centroid of the cluster is less than the threshold, then the sample belongs to the cluster; otherwise, the method creates a new cluster with this sample. The process applies to all samples of dataset. The first centroid is chosen as the first sample of the dataset. This scenario identifies the number of clusters dynamically. As a result, a new training dataset is built from the centroids of clusters. This model also employs SVM to classify the benign cases from malignant ones depending on the resultant training dataset from the first stage. At first in the preprocessing stage, the training dataset is separated into two categories - benign and malignant. Then, the modified K-means has been adopted to minimize the number of samples for each category with maintaining the high quality of data. The resultant high quality training dataset is then employed to train the SVM classifier with shortest training time, compared to when trained with full data, in addition to getting good results. Furthermore, the obtained new model is able to reach reliable results within minimal training time.

The rest of research has been organized as follows. Section 2 provides the state - of - the art breast cancer diagnosis methods, section 3 illustrates the methodology, section 4 discusses the obtained experimental results, and finally the main conclusions were listed in section 5.

2. Related work

The available literatures related to improving the performance of breast cancer diagnosis have been reviewed. The outcomes are presented in this section.

Sadhukhan et al. [18] analyzed digital image of a fine needle aspirate (FNA) of breast tissue and extract features of kernel of the cells and then compared KNN and SVM to predict breast cancer. The maximum accuracy which be obtained by applying KNN technique reach to 97.489%. Kumar et al. [19] proposed a hybrid fuzzy C-means combination with a cohort intelligence (CI) optimization algorithm to cluster data sufficiently and to deal with the expected limitations of the FCM. The Wisconsin breast cancer dataset was employed to check the validation of the hybrid methodology. The performance of proposed hybridized FCI such as accuracy did not clarify in the results of research. Wang et al. [20] proposed a convolutional neural network (CNN) which adopted on a modified Inception-v3 architecture to enable a good feature extraction of ABUS imaging. The developed CNN algorithm was tested and assessed using 316 breast cancer cases (135 malignant and 181 benign). They are achieved 94.68% AUC, 88.6% sensitivity and 87.6 specificity with 5-CV.

Stark et al. [21] developed six models of machine learning by applying Gaussian naïve Bayes, decision tree, discriminant analysis, logistic regression analysis, support vector machine, and feed-forward ANN. They yielded five-year breast cancer risk estimates that are more accurate than those resulted using only BCRAT tools. Therefore, they could improve the early detection and prevention of breast cancer. However, the outcome of neural network proved that the adopted machine learning model is

effective and efficient for both sets of input data as well as it confirmed that it can be more vigorous than the BCRAT.

AlFayez et al. [22], introduced a thermogram-based breast cancer detection approach. At first, a preprocessing of image was achieved by utilizing top-hat transform, homomorphic filtering in addition to an adaptive histogram equalization. Then, an implementation of binary masking and K-means algorithm were completed to segmentation ROI. Therefore, they used signature boundary for extraction of features. Finally, two classifiers were adopted and evaluated; these are Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM). However, the training time of proposed approach was not good for each of MLP and ELM. Ferroni et al. [23] presented an ML decision support system to obtain prognostic information from personal demographic, biochemical and clinical data concerning individuals with breast cancer - the extracted information were with adequately accurate (86%). Kyono et al. [24] proposed a DCNN based on screening cases to determine normal mammograms. They adopt a 10-fold cross validation to reveal that the DCNN has the ability to recognize 34% and 91% of the normal mammograms for a cancer prevalence of 15% and 1% respectively at 0.99 negative predictive value (NPV). Their work demonstrated the viability of employing DCNN to enhance radiologists' workflow productivity by not including the negative mammograms from reading, but the generalizability has yet to be validated with independent testing.

Omondiagbe et al. [25] examined artificial neural networks, support vector machine (utilizing a radial basis kernel), and Naïve Bayes based on the Wisconsin Diagnostic Breast Cancer data base. They focused on integrating these approaches in combination with techniques for feature selection and extraction in order to examine their performances to categorize the best adequate method. They concluded that a hybrid approach which reduced the high dimensionality of features by adopting discriminant analysis of linear type (LDA), and thereafter adopting the developed reduced feature dataset to support vector machine (SVM) had the ability to diagnosis breast cancer more precisely. Their proposed was able to achieve 98.82% accuracy, but they were used 70% of dataset for training and 30% for testing. Tapak et al. [26] carried out several works in order to investigate the possibility of diagnosing breast cancer accurately at early stages in addition to investigate in how to treat patients with metastases. They also tested the goodness of the two standard methods with the six algorithms related to machine learning. Consequently, the comparison revealed that the most acceptable act was obtained using the SVM method with accuracy level of 93% in the pathological grouping of invasive images for breast cancer. They were used 5-fold cross validation with 50% training and 50% validation of dataset.

Tseng et al. [27] employed the "serum human epidermal growth factor receptor 2" (sHER2) as logical features of clinic nature in order to expect the metastasis of breast cancer. In so doing, several machine learning techniques have been used such as SVM, random forest, Bayesian classification algorithms, and the statistical logistic regression method. Their analysis results confirmed that the random forest learning model was the most effective model to forecast the spread of breast cancer about 90 days in advance at least. The equivalent area below the curve of receiver operating characteristic was 0.75 with p-value < 0. 001.

Turkki et al. [28] proposed deep learning technique to predict of breast cancer by using tumor tissue images. The tissue microarray samples, extracted from a sample of 1299 patients with breast cancer, were taken nationwide. The samples were classified into groups according to their digital risk score (DRS) - low or high. An image sample for 868 patients were used to train the outcome classifier; the outputs were examined and compared with classification obtained from experts for a sample of 431 patients. Their research output pointed out the practicability of understanding predictive signals from the images of tumor tissue in the absence of domain knowledge. However, further validation is needed.

Xu et al. [29] proposed convolutional neural networks (CNNs) to segment three-dimensional ultrasound images of breast sample into four primary tissues: skin, mass, fatty tissue and fibro glandular tissue. They employed various quantitative measures for assessing segmentation outputs, these are accuracy, recall, precision and in addition to F1. The results of the developed CNNs method were with 80% accuracy, this implies the effectiveness and ability of this method in discriminating tissues developed in breast ultrasound images functionally. Furthermore, the value of the Jaccard index for similarity (JSI) was found to be as high as 85%.

3. Methodology

In this section, the proposed model that combines modified K-means with SVM is described in details. The methodology of the proposed model is demonstrated in Fig. 1.

At first, the datasets of breast cancer are read and prepared. The both adopted datasets of breast cancer,

Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC), have been employed to evaluate the performance of the proposed model. However, before training SVM, the dataset should be preprocessed by normalizing attributes with using Eq. (1).

$$X_{ij} = \frac{X_{ij} - min_j}{max_j - min_j} \quad (1)$$

Where
X_{ij} is a value of sample i in attribute j.
min_j is a minimum value of attribute j.
max_j is a maximum value of attribute j.

Moreover, a method to validate the proposed model is implemented by using 10-cross validation or by dividing the overall dataset into training and testing datasets based on ratio such as 75% for training and 25% for testing. Then, the training dataset is split into benign and malignant categories. The central aim of adopting the modified K-means was to decrease dataset size and also to create new training dataset of high quality and with small size. High quality implies that the samples of the original training dataset are completely represented in the resulting dataset. Based on such high-quality dataset, the SVM can result good prediction. There are several known adjusted K-means approaches adopted in previous studies; however, the modified K-means is the one that has recognized characteristics in that it can takes into account the entire set of probable eventualities by dealing with the whole divergent points within the dataset as preliminary centroids of clusters, instead of choosing a definite set of initial centroids arbitrarily, as is classically done. That is, modified K-means approach create clusters whereby the total cases are characterized by noteworthy differences among the samples. Therefore, based on the modified K-means, the dataset samples will be distributed to several suitable clusters with appropriate degree of accuracy. Nevertheless, the modified K-means differs from other K-means-based methods in that it does not require the computation of the number of clusters k because this will be achieved dynamically. The modified K-means is typically conducted on each single category with the purpose of minimizing the number of samples by categorizing them into clusters and then determining the mean for each cluster as a new sample. In the case of applying modified K-means in the benign category, for instance, the outcome would be a group of clusters consisted of similar samples. The new benign category samples, as a consequence, are quantified by finding the mean of each cluster and consider it as a new sample. The quality of the resultant samples should reflect the characteristics of the samples in the original training dataset. Fig. 2 shows the steps of modified K-means approach.

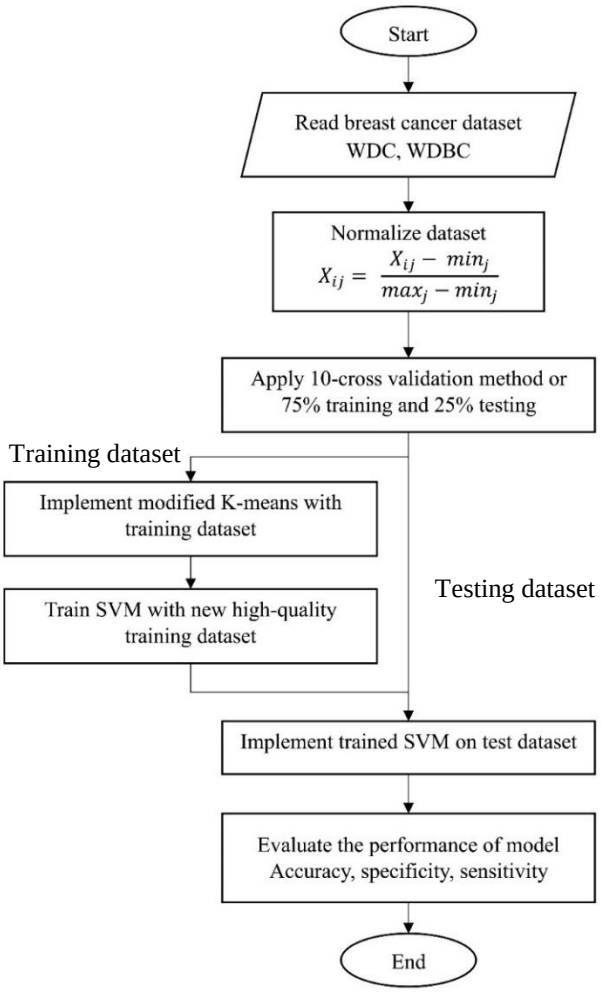

Figure. 1 The methodology of the proposed model

In the final step, the system utilizes the new obtained training dataset to inform the support vector machine classifier to depend upon various variables in enhancing the quality performance of the model developed for predicting breast cancer. The breast cancer classifier that is based on SVM is considered as a machine learning technique that utilizes statistical learning theories and techniques. Such classifier usually forms an instrument to divide data into various groups by an N-dimensional hyperplane which is determined based on a known training dataset. The dataset training samples are typically denoted as $\{(x_i, y_i)\}$, $i = 1, 2, ..., N$, where the number of samples is referred to by N and y_i represents the sample class x_i in the dataset. The central matter of the support vector machines is the finding of the maximum margin that isolates between hyperplane and the closest points in a high dimensional space. The SVMs determine the total distances between

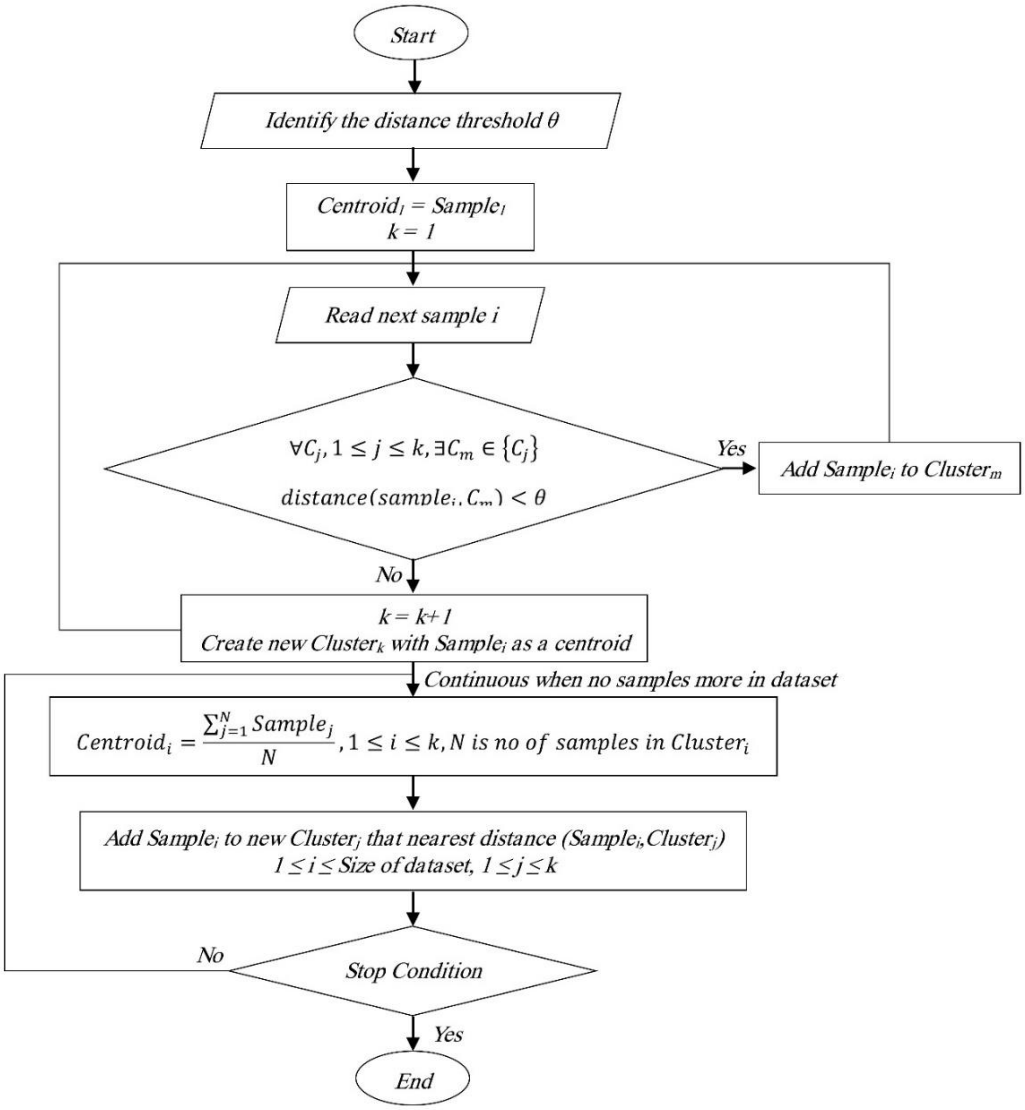

Figure. 2 The modified K-means algorithm

each point of hyperplane and the closest points of the space. The boundary function of the major margin is determined as follows, Eq. (2) [30].

$$Minimise\ W(\alpha)\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_i y_j \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^{N}\alpha_i \quad (2)$$

Subject to:

$$\forall i: 0 \le \alpha_i \le C\ and\ \sum_{i=1}^{N}\alpha_i y_i = 0$$

Where α is defined as a vector of N variables, C is the parameter that identifies the soft margin, $C > 0$, and $k(x_i, x_j)$ is the kernel function of SVMs. This group of kernel functions is beneficial in SVMs to divide the cases of data into diverse groups, such kernel functions can be mathematically specified as follows - Hsu et al., [30].

- Linear kernel: $k(x_i, x_j) = x_i^T . x_j$
- Polynomial kernel: $k(x_i, x_j) = (\gamma x_i^T . x_j + r)^d, \gamma > 0$
- Radial basis function (RBF) kernel: $k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0$
- Sigmoid kernel: $k(x_i, x_j) = \tanh(\gamma x_i^T . x_j + r)$

Where γ, r and d are kernel parameters.

The main goal of each proposed classification model is reducing bias and variance issues in classification. Bias can be described as a systematic error of a specific learning method and it is usually influenced by the learning algorithm or method itself [31]. Variances reflect and identify the random errors, which are happened due to the uncertainty in learning method procedure or in training data. Speaking broadly, prediction models with low bias can lead to

overfitting problems which in turn can compromise the model accuracy and even the ability to better classify new dataset samples. In contrast, developed models with low variance can lead to underfitting issues and as a consequence ending with inaccurate outcomes. A single SVM, as one of the well-known techniques that take into account both variance and bias likely issues when conducting the cancer classification process. The variety and number of setting parameters in the developed models of SVM can greatly influence the precision of classification. In specific, the accuracy of breast cancer classification depends to large extent on the diverse alternatives of kernel functions and also on the structure of the SVMs. In the current proposed model, a radial basis function is used and SVM with grid search algorithm are adopted to choose the best parameters for SVM.

In the testing phase, the trained SVM with high quality training dataset is used to classify the testing dataset into benign and malignant categories. One of the frequently adopted methods in literature to quantify and assess the performance quality for the classification models is the confusion matrix. In the confusion matrix, the examined breast cancer cases are divided into two distinct classes: positive (Benign) and negative (Malignant). Table 1 illustrates how predicted and actual classes (cases) are compared with each other to yield four distinguishing metrics:

- True Positives (TP) – indicate positive cases that are correctly diagnosed as positive cases.
- False Positives (FP) – indicate negative cases that are incorrectly diagnosed as positive cases.
- True Negatives (TN) – indicate negative cases that are correctly diagnosed as negative cases.
- False negatives (FN) – indicate positive cases that are incorrectly diagnosed as negative cases.

Based on the confusion matrix, other performance metrics can be derived as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Sensitivity = Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$F-Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (7)$$

Table 1. Confusion matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	TP	FN
	Negative	FP	TN

Nevertheless, the TP, FP, TN, and FN cases can be gathered and arranged to sketch a plot that is known as a Receiver Operating Characteristic (ROC) curve. The ROC curve can aid in manifesting the negative impacts of the number of FP and FN cases on the model classification errors. In addition, depending upon the ROC sketch, the area under curve (AUC) can be determined. Let the values of α and $1-\beta$ represent the likelihoods of TP and FP respectively. Then, the needed area is predicted by the trapezoidal technique of integration – as stated in Eq. (8) below [32]:

$$AUC = \sum_i \{[\alpha_i.\Delta(1-\beta)] + \frac{1}{2}[\Delta\alpha.\Delta(1-\beta)]\} \quad (8)$$

Where

$$\Delta(1-\beta) = (1-\beta)_i - (1-\beta)_{i-1}$$
$$\Delta\alpha = \alpha_i - \alpha_{i-1}$$

4. Experimental results

This section is devoted to appraise the prediction goodness of the new developed modified K-means model and the classification capability of the SVM algorithm. The entire experimental and modelling work were carried out on a PC with Windows-10 operating system, Intel Core i5 CPU@2.60 GHz and with RAM of 12 GB. MATLAB (ver. 2019) and c-SVC with RBF kernel of LIBSVM (ver. 3.24) have been utilized to build the required codes for all the implementations.

4.1 Datasets

The effectiveness of the proposed model was evaluated by using WBC (Wisconsin Breast Cancer) and WDBC (Wisconsin Diagnostic Breast Cancer) datasets which are two typical breast cancer datasets. The two recognized datasets were previously gathered by Mangasarian et al. [33] from the University of Wisconsin Hospitals. Each sample was labelled as either benign or malignant. The WBC dataset consists of 699 cases (approximately 65.5% benign and 34.5% malignant). In addition, it includes 11 patient characteristics; these are patient ID, nine key features for tumor and also one class indicator. Through data screening and cleaning, 16 samples with missing attributes were excluded. In contrast,

the WDBC consists of 569 cases (approximately 62.7% benign and 37.3% malignant) in addition to 32 key patient attributes. These attributes are patient ID, 30 tumor distinguishing features, and one class indicator. The distinguishing features of patients' tumors were gathered based on 10 different aspects: texture, radius, area, perimeter, smoothness, concavity, compactness, concave points, fractal dimension and symmetry. These features were obtained using digitized image of a fine needle aspirate (FNA) of a breast mass. Key identifying statistics for each image such as mean, standard error, and smallest or largest values of these features were determined and hence leading to a set of 30 features.

4.2 Experiments

The first experiment shows the ability of modified K-means to minimize the number of samples of datasets. Table 2 shows the number of samples of the WBC and WDBC datasets before and after applying modified K-means with distance threshold of 0.9 and 0.8, respectively. The quality of the resultant samples represents all of the samples in the original datasets. Moreover, the performance of proposed model in term of accuracy was improved by using modified K-means.

In each evaluative test, the specificity, accuracy, sensitivity, precision, F-score and AUC for WBC and WDBC datasets have been computed. The performance of the proposed modified K-means and SVM model is better than the traditional single SVM classifier. Accordingly, the proposed model can better classify the breast cancer tumors than a single model as shown in Table 2. The overall results of WBC and WDBC are shown in Tables 3 and 4, respectively. In these experiments, the size of training dataset is 75% of the original dataset. The largest values have been distinguished in bold for each performance metric in each table. The needed training times are shown in the last column of each table.

The comparison results of proposed model with other classifiers by Weka tool (Bayes Network, Random Forest, K-Nearest Neighbor and Naïve Bayes) demonstrate that the new model behaver better than others with respect to accuracy and other performance metrics as shown in Table 3 and Table 4.

Table 2. Number of dataset samples before and after applied modified K-means

Dataset	No. of samples		Reduction ratio	Accuracy	
	Before	After		Single SVM	Proposed model
WBC	524	36	93%	97.714	**98.286**
WDBC	426	31	93%	97.902	**98.601**

Table 3. Comparison of proposed model with other methods based on 75% of training samples on WBC dataset

Method	Accuracy	Sensitivity	Specificity	Precision	F-Score	AUC	Training time (ms)
Single SVM	97.714	97.479	98.214	**99.145**	98.305	96.511	0.0055
Bayes Net	96.571	96.364	96.923	98.148	97.248	95.6	-
Naïve Bayes	94.857	94.545	95.385	97.196	95.852	94.7	-
K-nearest	96	96.364	95.385	97.248	96.804	96.1	-
Random Forest	96	96.364	95.385	97.248	96.804	97.7	-
Proposed model	**98.286**	**98.165**	**98.485**	99.074	**98.617**	**98.325**	**0.0025**

Table 4. Comparison of proposed model with other methods based on 75% of training samples on WDBC dataset

Method	Accuracy	Sensitivity	Specificity	Precision	F-Score	AUC	Training time (ms)
Single SVM	97.902	**100**	93.333	97.03	98.493	96.667	0.0053
Bayes Net	93.662	94.186	92.857	95.294	94.737	92.9	-
Naïve Bayes	91.549	91.86	91.071	94.048	92.941	92.3	-
K-nearest	95.775	97.674	92.857	95.455	96.552	94.6	-
Random Forest	96.479	98.837	92.857	95.506	97.143	96.1	-
Proposed model	**98.601**	98.98	**97.778**	**98.98**	**98.98**	**98.379**	**0.0023**

The best average accuracy achieved by the modified K-means and SVM model is 98.286% on the WBC dataset and 98.601% on the WDBC dataset.

To compare the performance of the proposed modified K-means and SVM model with other similar methods used in previous research, several existing algorithms were conducted as a benchmark for both WBC and WDBC datasets. Ten-fold cross-checking method was employed to examine the effectiveness of the existing models. The original accuracy results for these models based on previous studies have been listed in Tables 5 and Table 6. For the WBC dataset, Table 5 shows that the proposed model is quite competitive and outperforms most of the other classifiers with average reduction ratio reach to 92%. However, the performance in terms of specificity and precision for Bayes Network is slightly better. In terms of accuracy, the developed model is the one with the best accuracy. Table 6 depicts the comparison measures for the WDBC cancer dataset with average reduction ratio reach to 91%.

Based on the evaluative measures listed in the preceding tables, it appears evident that the current developed classification model is quite better than the others. In this research, the merits of the proposed K-means and SVM model outweigh those for the implemented SVM-RBF kernel model whereby cross-validation of 10-fold type was carried out. Furthermore, the new developed classification model was also able to outperform the model previously developed by [35], which depends on utilizing the feature selection process. The above results show that the proposed model (modified K-means and SVM) can enhance the performance of breast cancer diagnosis. The proposed model showed higher reliability than all other currently used models and it is worthwhile noting that it has only minor consequences on the stability of predictions. Furthermore, it is important to emphasize the strength of the proposed model whereby its accuracy outweighs the other considered ones. Finally, the comparison revealed how the developed model requires comparatively less training time which is a pure result of the effective reduction in the sample size of the original training dataset.

5. Conclusion and future work

The current research presents a new breast cancer prediction model that is based on both modified K-means and support vector machine algorithm. The new developed model has obvious better classification performance than the single SVM one. The modified K-means has been considered as a means to attain high-quality training datasets where the reduced training time can effectively improve the entire performance of the adopted support vector machine. Based on the analytic results obtained from the two experimental training datasets - WBC and WDBC, the new developed model has attained 96.996% and 98.067% degrees of accuracy respectively when used 10-cross validation method. Furthermore, in terms of sensitivity, specificity and AUC our proposed can achieve up to 97.16%, 96.68% and 96.92% with WBC dataset respectively, while 100%, 94.81% and 97.41% with WDBC dataset respectively. The major achievement of the current research is the developing of a prediction model with better performance than those models reported in recent relevant works. This is because the developed model can present a balanced performance among categories malignant and benign. Based on the findings obtained, it is recommended to develop more robust model based on competent classifiers as an ensemble to proficiently classify breast cancer cases with high performance.

Table 5. Comparison of proposed model with other methods based on 10-CV on WBC dataset

Method	Accuracy	Sensitivity	Specificity	Precision	F-Score	AUC
Single SVM	96.567	96.943	95.851	97.797	97.368	96.397
Bayes Net	96.853	96.07	**98.34**	**99.099**	97.561	95.122
Naïve Bayes	95.994	95.197	97.51	98.643	96.889	93.8
K-nearest	95.136	96.725	92.116	95.887	96.304	91.5
Random Forest	96.71	96.943	96.266	98.013	97.475	96.5
CNN [20]	-	88.6	87.6	-	-	94.68
SVM [26]	93	-	-	-	-	-
CNN [29]	80	-	-	-	-	-
Proposed model	**96.996**	**97.162**	96.68	98.234	**97.695**	**96.921**

Table 6. Comparison of proposed model with other methods based on 10-CV on WDBC dataset

Method	Accuracy	Sensitivity	Specificity	Precision	F-Score	AUC
Single SVM	97.54	99.44	94.34	96.73	98.066	96.89
Bayes Net	95.079	96.078	93.396	96.078	96.078	95.1
Naïve Bayes	92.97	94.958	89.623	93.906	94.429	92.8
K-nearest	95.958	96.919	94.34	96.648	96.783	95.5
Random Forest	96.309	98.039	93.396	96.154	97.087	96.35
Fuzzy Clustering [34]	95.57	-	-	-	-	-
Weighted vote-based ensemble [35]	95.09	98.60	-	94.10	96.297	-
Proposed model	**98.067**	**100**	**94.811**	**97.011**	**98.483**	**97.406**

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, software, investigation, validation, writing—original draft preparation, writing and editing, Wathiq Laftah Al-Yaseen; resources, Wathiq Laftah Al-Yaseen and Ammar Jehad; review, Wathiq Laftah Al-Yaseen, Qusay Abdullah Abed and Ali Kadhum Idrees.

References

[1] L. G. Ahmad, A. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence", *Journal of Health and Medical Informatics*, Vol. 4, No. 2, pp. 1-3, 2013.

[2] F. Liu and M. Brown, "Breast Cancer Recognition by Support Vector Machine Combined with Daubechies Wavelet Transform and Principal Component Analysis", In: *Proc. of the International Conf. on ISMAC in Computational Vision and Bio-Engineering*, Springer, pp. 1921-1930, 2018.

[3] L. Yang and Z. Xu, "Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning", *International Journal of Machine Learning and Cybernetics*, Vol. 10, No. 3, pp. 591-601, 2019.

[4] M. Alkhaleefah and C.-C. Wu, "A hybrid CNN and RBF-based SVM approach for breast cancer classification in mammograms", In: *Proc. of 2018 IEEE International Conf. on Systems, Man, and Cybernetics (SMC)*, pp. 894-899, 2018.

[5] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform", *Measurement*, Vol. 146, pp. 800-805, 2019.

[6] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", *Journal of Algorithms and Computational Technology*, Vol. 12, No. 2, pp. 119-126, 2018.

[7] E. Alickovic and A. Subasi, "Normalized neural networks for breast cancer classification", In: *Proc. of International Conf. on Medical and Biological Engineering*, Springer, pp. 519-524, 2019.

[8] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction", *Radiology*, Vol. 292, No. 1, pp. 60-66, 2019.

[9] J. Holm, J. Li, H. Darabi, M. Eklund, M. Eriksson, K. Humphreys, P. Hall, and K. Czene, "Associations of breast cancer risk prediction tools with tumor characteristics and metastasis", *Journal of Clinical Oncology*, Vol. 34, No. 3, pp.251-258, 2020.

[10] Y. S. Vang, Z. Chen, and X. Xie, "Deep learning framework for multi-class breast cancer histology image classification", In: *Proc. of International Conf. Image Analysis and Recognition*, Springer, pp. 914-922, 2018.

[11] H. M. Moftah, A. T. Azar, E. T. Al-Shammari, N. I. Ghali, A. E. Hassanien, and M. Shoman, "Adaptive k-means clustering algorithm for MR breast image segmentation", *Neural Computing and Applications*, Vol. 24, No. 7-8, pp. 1917-1928, 2014.

[12] A. K. Dubey, U. Gupta, and S. Jain, "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset", *International Journal of Computer Assisted Radiology and Surgery*, Vol. 11, No. 11, pp. 2033-2047, 2016.

[13] B. C. Patel and G. Sinha, "An adaptive k-means clustering algorithm for breast image

segmentation", *International Journal of Computer Applications*, Vol. 10, No. 4, pp. 35-38, 2010.

[14] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Automatic breast cancer diagnosis based on k-means clustering and adaptive thresholding hybrid segmentation", *Image processing and communications challenges*, Springer, pp. 295-302, 2011.

[15] N. Singh, A. G. Mohapatra, and G. Kanungo, "Breast cancer mass detection in mammograms using K-means and fuzzy C-means clustering", *International Journal of Computer Applications*, Vol. 22, No. 2, pp. 15-21, 2011.

[16] A. K. Dubey, U. Gupta, and S. Jain, "Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data", *International Journal on Advanced Science, Engineering and Information Technology*, Vol. 8, No. 1, pp. 18-29, 2018.

[17] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system", *Expert Systems with Applications*, Vol. 67, pp. 296-303, 2017.

[18] S. Sadhukhan, N. Upadhyay, and P. Chakraborty, "Breast Cancer Diagnosis Using Image Processing and Machine Learning", *Emerging Technology in Modelling and Graphics*, Springer, pp. 113-127, 2020.

[19] M. Kumar, A. J. Kulkarni, and S. C. Satapathy, "A Hybridized Data Clustering for Breast Cancer Prognosis and Risk Exposure Using Fuzzy C-means and Cohort Intelligence", *Optimization in Machine Learning and Applications*, Springer, pp. 113-126, 2020.

[20] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning", *Ultrasound in Medicine and Biology*, Vol. 46, No. 5, pp. 1119-1132, 2020.

[21] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting breast cancer risk using personal health data and machine learning models", *Plos One*, Vol. 14, No. 12, pp. 1-17, 2019.

[22] F. AlFayez, M. W. A. El-Soud, and T. Gaber, "Thermogram Breast Cancer Detection: a comparative study of two machine learning techniques", *Applied Sciences*, Vol. 10, No. 551, pp. 1-20, 2020.

[23] P. Ferroni, F. M. Zanzotto, S. Riondino, N. Scarpato, F. Guadagni, and M. Roselli, "Breast cancer prognosis using a machine learning approach", *Cancers*, Vol. 11, No. 328, pp. 1-9, 2019.

[24] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Improving workflow efficiency for mammography using machine learning", *Journal of the American College of Radiology*, Vol. 17, No. 1, pp. 56-63, 2020.

[25] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis", In: *Proc. of IOP Conf. Series: Materials Science and Engineering*, Vol. 495, pp. 1-16, 2019.

[26] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers", *Clinical Epidemiology and Global Health*, Vol. 7, No. 3, pp. 293-299, 2019.

[27] Y. J. Tseng, C. E. Huang, C. N. Wen, P. Y. Lai, M. H. Wu, Y. C. Sun, H. Y. Wang, and J. J. Lu, "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies", *International Journal of Medical Informatics*, Vol. 128, pp. 79-86, 2019.

[28] R. Turkki, D. Byckhov, M. Lundin, J. Isola, S. Nordling, P. E. Kovanen, C. Verrill, K. von Smitten, H. Joensuu, J. Lundin, and N. Linder, "Breast cancer outcome prediction with tumour tissue images and machine learning", *Breast Cancer Research and Treatment*, Vol. 177, pp. 41-52, 2019.

[29] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning", *Ultrasonics*, Vol. 91, pp. 1-9, 2019.

[30] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification", *Computer Science*, pp. 1-16, 2008.

[31] A. Rosales-Pérez, H. J. Escalante, J. A. Gonzalez, C. A. Reyes-Garcia, and C. A. C. Coello, "Bias and variance multi-objective optimization for support vector machines model selection", In: *Proc. of Iberian Conf. on Pattern Recognition and Image Analysis*, Springer, LNCS 7887, pp. 108-116, 2013.

[32] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, Vol. 30, No. 7, pp. 1145-1159, 1997.

[33] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", *Operations Research*, Vol. 43, No. 4, pp. 570-577, 1995.

[34] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers", *Pattern Recognition Letters*, Vol. 24, No. 14, pp. 2195-2207, 2003.
[35] S. Bashir, U. Qamar, and F. H. Khan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble", *Quality and Quantity*, Vol. 49, No. 5, pp. 2061-2076, 2015.