Computer Systems Science & Engineering
DOI:10.32604/csse.2022.022942
Article

Tech Science Press

Diagnosing Breast Cancer Accurately Based on Weighting of Heterogeneous Classification Sub-Models

Majdy Mohamed Eltayeb Eltahir[1,*] and Tarig Mohammed Ahmed[2,3]

[1]Department of Information Systems, College of Science & Arts at Mahayil, King Khalid University, Muhayel Aseer, 62529, Kingdom of Saudi Arabia
[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, 21589, Kingdom of Saudi Arabia
[3]Department of Computer Sciences, Faculty of Mathematical Sciences and Informatics, University of Khartoum, Khartoum, 11115, Sudan
*Corresponding Author: Majdy Mohamed Eltayeb Eltahir. Email: meltahir@kku.edu.sa
Received: 23 August 2021; Accepted: 20 October 2021

Abstract: In developed and developing countries, breast cancer is one of the leading forms of cancer affecting women alike. As a consequence of growing life expectancy, increasing urbanization and embracing Western lifestyles, the high prevalence of this cancer is noted in the developed world. This paper aims to develop a novel model that diagnoses Breast Cancer by using heterogeneous datasets. The model can work as a strong decision support system to help doctors to make the right decision in diagnosing breast cancer patients. The proposed model is based on three datasets to develop three sub-models. Each sub-model works independently. The final diagnosis decision is taken by the three sub-models independently. The power of the model comes from the diversity checks of patients and this reduces the risk of wrong diagnosing. The model has been developed by conducting intensive experiments. Several classification algorithms were used to select the best one in each sub-model. As the final results, the sub-model accuracies were 72%, 74% and 97%.

Keywords: Breast cancer; data mining; classification

1 Introduction

In the developed and developing countries, breast cancer is one of the leading forms of cancer affecting women alike. As a consequence of growing life expectancy, increasing urbanization and embracing Western lifestyles, the high prevalence of this cancer is noted in the developed world [1]. Although prevention interventions may help reduce some of the risks of developing breast cancer, they cannot eradicate most cases of breast cancer that arise in low- and middle-income countries where it is not detected until the late stages of the disease. Therefore, the key to treating this disease remains early diagnosis to increase breast cancer outcomes and improve patient survival rates [2]. Education and awareness of breast cancer's early signs and symptoms will be life-saving. The variety of therapies available is broader and more diverse as the illness is diagnosed in its early initial stages, and the odds of a complete recovery are very

great. Many of the lumps that are found in the breast are not malignant, but the presence of a lump or thickening of the breast tissue is the most common early symptom of breast cancer in women and men alike. Usually, this lump is painless.

This paper aims to develop a novel model that diagnoses Breast Cancer by using heterogeneous datasets. The model can work as a strong decision support system to help doctors to make the right decision in diagnosing breast cancer patients. The proposed model is based on three datasets to develop tree sub Clustering-models. Each sub-model works independently. The final diagnosis decision is taken by the three sub-models independently. The power of the model comes from the diversity checks of patients and this reduces the risk of wrong diagnosing. The model has been developed by conducting intensive experiments. Several classification algorithms were used to select the best one in each sub-model. Most of the issued models were based on only one dataset which made the diagnosing is not accurate. The obtained results have been evaluated and discussed in details.

The paper rest is organized into four sections: section two presents SVM and Naïve Bayes algorithms. These algorithms were used to develop the model. Section three presents some interesting researches. Section four describes the research model in terms of framework, datasets, implementation and discussion. Finlay, section five concludes the paper with some recommendations as future works to enhance the model.

2 Data Mining Algorithms

To develop an efficient data mining model, the appropriate data mining algorithm plays a crucial role. Data mining algorithms are classified into two groups: supervised and unsupervised. This research focuses on the supervised group to classify breast cancer patients. The following algorithms have been used in this research to develop our proposed model:

2.1 SVM

Support Vector Machines is a machine learning algorithm. It is a directed learning algorithm. It can be used for either regression or classification task [3]. It is based on the idea of creating a hyperplane that divides the data set into two classes in the best way, as shown in Fig. 1. The hyperplane is the line that separates linearly and classifies a dataset. In each of the data sets, the margin is the distance between the hyper plane and the nearest point. The aim is to choose a super-level that has the largest margin between it and any point in the training data set in order to maximize the probability of new data being correctly classified.

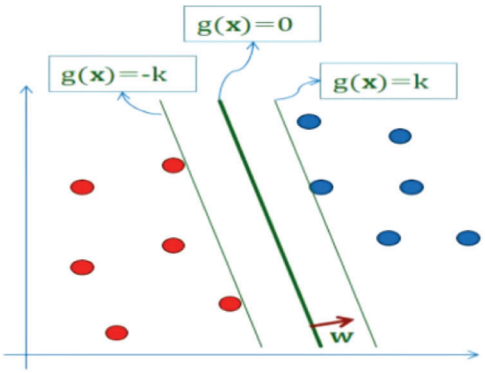

Figure 1: SVM

$$g(x) = w^T + b \tag{1}$$

Maximize k such that:

$$-w^T x + b \geq \text{ k for } d_i == 1$$

$$-w^T x + b \leq \text{ k for } d_i == -1$$

Value of g(x) depends only upon ||w||:

1) Keep w = 1 and maximize g(x) or,
2) g(x) > 1 and minimize ||w||

Approach 2 is used, and the problem is stated as follows:

$$\varnothing(w) = \frac{1}{2} w^T w - minimize \tag{2}$$

Subject to $d_i\,(w^T x + b) \geq 1 \forall i$

$$Minimize: J(w,\ b,\ \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i (w^T x_I + b) + \sum_{i=1}^{N} \alpha_i \tag{3}$$

Subject to $\alpha_i \geq 0 \forall i$

According to the Lagrange multiplier method, J is minimized for w and b, but it must be maximized for **α**. Method of Lagrange multipliers states that J is minimized for w and b as before, but it has to be **maximized** for **α**. We can transform the function J into its dual form for the solution since it is currently expressed in its primal form.

$$J(w,\ b,\ \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i (w^T x_I + b) + \tag{4}$$

At the optimum $\frac{\delta j}{\delta w} = 0$ and $\frac{\delta j}{\delta b} = 0$

So we can write J as:

$$J(w,\ b,\ \alpha) = \sum_{i=1}^{N} \alpha_i + \frac{1}{2} w^T w - w^T \sum_{i=1}^{N} \alpha_i d_i x_i - b \sum_{i=1}^{N} \alpha_i \tag{5}$$

Then

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T \tag{6}$$

Q(α) represents the dual form J which is only dependent on α as rest are all known scalars.

2.2 Naïve Bayes

The Naïve Bayes classifier is based on Bayes' theory of probability [4].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

The form of training data in the Bayes Classifier is of the form: {(Xi, Wi)} Where Xi is a set of Attributes to an element and sometimes called Attributes Vector, or the Features Vector, of shape

$$X = [x_1 x_2, \ \ldots, \ x_d] \tag{8}$$

Wi is the category to which the element belongs, where the element is classified based on the attributes values.

The Naïve bay assumes independence between the Attributes, the Attributes have no relationship, as they do not affect each other Mathematically like this

$$P\ (x1, \ \ldots, \ x_d|\omega) = \prod_i P(x_I|\omega_I) \tag{9}$$

In the Naïve Bayes Classifier, the element is classified into the group in which the probability that it is the greatest probability (arg max) is simply to have a new element:

$$\omega_{NB} = \arg max\ P(\omega_j) \prod_i P(x_I|\omega_I) \tag{10}$$

3 Related Work

As mentioned above, breast cancer is considered as one of the most serious diseases that cause death among women. The researchers have conducted many types of research related to this area. These researches were deal with breast cancer in terms of diagnosing and treatments. In this section presents some of these researches as following:

Etehadtavakol et al. [5] proposed a model by using data mining algorithms. The model used a dataset from Breast Cancer Surveillance Consortium (BCSC). Seven risk factors were modelled from the data set. Clustering and correlation were applied to discover rising breast cancer.

The accuracy of the model after implementation was 89.6%. Mohanty et al. [6] used classification techniques to help in detecting breast cancer. They proposed a system that was working based on three steps: (1) extraction of the region of interest (ROI) with size 256 × 256. (2) Extracting 26 features and using the features to detect breast cancer. (3) Using association rule to classify breast cancer cases or not. The system used the digital mammogram. A computer system proposed by Kuo et al. [7] that used a data mining classification technique (Decision Tree) to classify a tumour of the breast. The proposed system had worked as an assistant to physicians to diagnose the disease. The system model used a Region of Interest (RIO) image to extract the case feature as input information. The decision tree algorithm classifies the cases. The system mentioned that can classify breast cancer with 86.67% as accuracy rate.

Two algorithms were used to build a model proposed by Mousa et al. [8] the algorithms: Wavelet and Fuzzy-neural worked as a classifier of mammographic. The model can detect benign or malignant. It used a dataset from Mammography Image Analysis Society (MIAS). The accuracy of the model was accepted. A comparative study was conducted by Diz et al. [9] to compare two datasets of breast cancer. The study aimed to select the best predictive methods that can diagnose breast cancer. To achieve its goal, the study used Matlab and classification algorithm on Weka data mining application. As result, the accuracy was found between 89.3% to 64.7%. Naïve Bayes algorithm was selected as the best one. Santos et al. [10] developed several models that can detect breast cancer in early stages by using X-ray images. Many classification algorithms had been used. The models were evaluated using AREA UNDER THE CURVE (AUC) and confusion matrix. Naïve Bayes was select as the best algorithm based on performance output.

By using a feature selection method “INTERACT”, Shen et al. proposed a classification model to diagnose breast cancer. The accuracy of the model was improved when using the feature selection method. The model was built based on 9 selected factors that relevant to a breast cancer diagnosis. The selection process aimed to improve the quality of the model [11]. Mansour et al. [12] developed a model

based on a massive dataset related examined genes. The data was generated from the microarray. The model used a clustering algorithm to detect breast cancer groups. In addition, the model applied 24,481 genes of breast tumour samples. This study found that 17 genes had a relation with breast cancer marker genes.

By using thermography, breast cancer could be detected in early stages. Mookiah et al. [13] developed a model that used 50 thermograms to classify normal and abnormal groups. The model also used WAVELET and texture as feature selection to enhance the model quality. Multiple data mining algorithms were used. The highest accuracy average was 93.3% and the sensitivity was 86.7%. Based on more than 200,000 as a dataset, Delen et al. [14] developed a data mining model. As algorithms, the model used a logistic regression tree, a decision tree, and neural network. Performance analysis had been conducted between the three models. According to the results, the decision tree was reported as the best model (93.6% accuracy). The second was a neural network (91.2% accuracy). The third was Logistic Regression (89.2% accuracy). By using SEER dataset which contains 151,886 records, Sarvestani et al. [15] proposed a model that can predict breast cancer survival patients. The model used C4.5 algorithms, neural network and Naïve Bayes. In addition, 16 attributed were selected to represent each patient. The C4.5 algorithm was selected as the best one with high performance.

A reduced set of discriminatory characteristics from curvelet transform for the diagnosis of breast cancer was addressed by Dhahbi et al. [16] and suggested a feature extraction method depended on curvelet transform and mammogram kind of moment theory description. Their Research findings show the effectiveness and superiority of mammogram study curvelet moments. Indeed, the mini-MIAS database results show the curvelet moment yield accuracy of 91.27 per cent (respectively 81.35 per cent) with 10 (resp.8) abnormality (resp. malignancy) detection features. Turgut et al. [17] used data on microarrays of breast cancer for patient selection through the application of machine learning methods. Firstly, two distinct methods of feature selection were applied to the data by 8 separate algorithms for learning machines. The techniques used are SVM, MLP, KNN, Random Forest, Decision Trees, Logistic Regression, Gradient Boosting Machines and Ad boost. SVM delivered the best results after implementing the two different methods of selecting apps with the best 50 apps. MLP is implemented with different layers and neurons to test the effect of the number of layers and neurons on classification accuracy. It is determined that the number of layers has frequently decreased, and the consistency has not improved sometimes. To overcome these forms of conditions, Bala et al. [18] have examined breast tissue. Better predictive methodologies for diseases are supported by machine learning in managing health care. Ensemble learning is nothing more than a sequence of classifiers that actually generate better outcomes than the existing ones. In order to achieve the best performance, they use arrays of classifiers known as ensembles. Along with current classifiers such as J48Naive Bayes, Random Forest and SMO, they have introduced ensemble strategies to enhance the better estimation of breast cancer to identify breast tissue as in the type of carcinoma and fibroadenoma. Ensemble classifiers such as Ada boosting, bagging and piling or mixing strategies with them were also added. A detailed comparative study of different classification strategies used to predict breast cancer is proposed by Abd-Elrazek et al. [19]. With a set of traditional supervised machine learning and data mining techniques, predicting breast cancer (benign or malignant) can be used. They suggest and apply three different algorithms to two different sets of data. These algorithms are Classification without Feature Selection (CWFS), Feature Selection Classification (FSC), and Standardization and Feature Selection Classification (NFSC). The outcomes of accuracy, specificity, precision and sensitivity are calculated and recorded for each device. Therefore, their observations suggest greater accuracy relative to up-to-date techniques Balaraman et al. [20] made an attempt with the Naive Bayes, Multilayer Perceptron, Radial Base Function Network, nearest neighbor, Conjunctive rule to boost the efficiency of detection, data set accuracy for breast cancer. Basic train-test data ratio checks were carried out and an average accuracy of over 98 percentage points is obtained by this hybrid classifier. Comparative studies were carried out for this analysis using various

supervised machine-learning techniques that are useful for forecasting the rise in breast cancer rates, while previous researchers used various analyzes, such as linear regression, Random Forest, Multilayer perceptron, Decision Tree, for different approaches to machine learning [21]. The behavior and conclusions of a systematic review (SR) aimed at reviewing the state of the art with regard to computer-aided diagnosis/detection of breast cancer (CAD) systems is presented by Yassin et al. and Segera et al. [22,23] In the classification of cancer-based on imbalances in microarrays, an important but difficult combinatorial function is decided to be an optimal decision model. Although a major contribution has already been made in this field by the Multiclass Support Vector Machine (MCSVM), its performance depends solely on three aspects: the penalty factor C, the form of the kernel and its parameters.

The conventional diagnosis of breast cancer also raises challenges such as poor levels of accuracy and limited self-adaptability. In order to solve these problems, the author has proposed an Ada Boost-SVM classification algorithm, combined with k-means in this work for early breast cancer diagnosis. Through measuring its precision, the uncertainty matrix that provides doctors with valuable hints for the detection of early breast cancer, the useful nesses of the suggested approaches were tested [24]. Machine learning algorithms are used to classify benign and malignant tumors in which computers learn from previous data and the most current input category can be expected. A performance distinction is made in this analysis on the Wisconsin Breast Cancer datasets from the data sets of the UCI Machine Learning Repository between different support vector machine (SVM) machine learning algorithms, artificial neural network (ANN), Bayesian Network (NB) and k-nearest neighbor algorithms (K-NN). The experimental findings showed that in the role of identifying breast cancer, supporting vector machine techniques obtained greater accuracy, achieving precision within the range of 97.6 per cent to 98.8 per cent, which is a remarkable accuracy for controlled classification of patterns [25]. Gupta et al. [26] noted that the precision of the diagnostic analysis of the various methods used to identify data mining is highly acceptable and can help medical professionals make early diagnostic decisions and avoid biopsies. The topic is studied specifically in the sense of Artificial Neural Networks (ANNs), which, relative to other approaches used for classification, resulting in high precision. The authors propose that the optimal model can be achieved by designing several different types of models by using different techniques and algorithms. In their report, PadmaPriya et al. [27] explored that breast cancer recurrence has for many years been the most demanding of researchers. The precise cause of breast cancer is uncertain, but early detection may be a better way of avoiding and identifying breast cancer. They concluded that more reliable results are given by data mining algorithms such as Decision Tree, Naïve Bayes and Help Vector Machine.

4 Research Model

In this paper, a novel model has been introduced to diagnose the barest cancer effectively based on the heterogeneous sub-models. The main idea behind this model is to use heterogeneous datasets to develop multiple classification sub-models. Each one of sub-model participates in diagnosing breast cancer patients in a different way. As we mentioned above, breast cancer is a very dangerous disease. Therefore, the diagnosing accuracy is a critical issue which reflects on patients' lives. To develop an effective model to diagnose breast cancer, we use heterogeneous medical record datasets. Each dataset consists of multiple features to represent patients' medical records in a different way. Not like the current breast cancer classification models, the proposed model can give more accurate results because it is based in heterogeneous datasets. The research model uses five datasets to develop five sub-models. Each sub-model participates in diagnosing the patients by using a weighting method. The weighting method represents the percentage of the model in tacking a decision of the diagnosing. Fig. 2 presents the model framework.

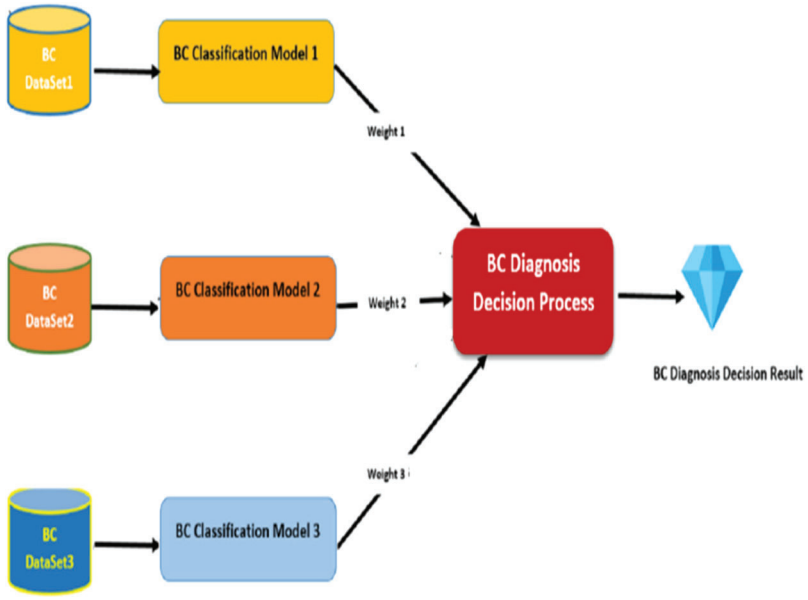

Figure 2: Proposed model framework

4.1 Model Datasets

The research model uses five heterogeneous datasets. Each one is used to develop a sub-model. In this section, the datasets are described as following:

4.1.1 Dataset 1

Dataset1 consists of 286 instances and 9 attributes plus the class, some of which are nominal and some are linear. It is created by Matjaz Zwitter & Milan Soklic (physicians), Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia [28]. The following Tab. 1 and Fig. 3 describe the dataset.

Table 1: Dataset 1 attributes description

Attribute with description
1. Class (no-recurrence-events, recurrence-events)
2. Age
3. Menopause (lt40, ge40, premeno)
4. Tumor-size
5. Inv-nodes
6. Node-caps (yes, no)
7. Deg-malig (1, 2, 3)
8. Breast (left, right)
9. Breast-quad (left-up, left-low, right-up, right-low, central)
10. Irradiat: yes, no.

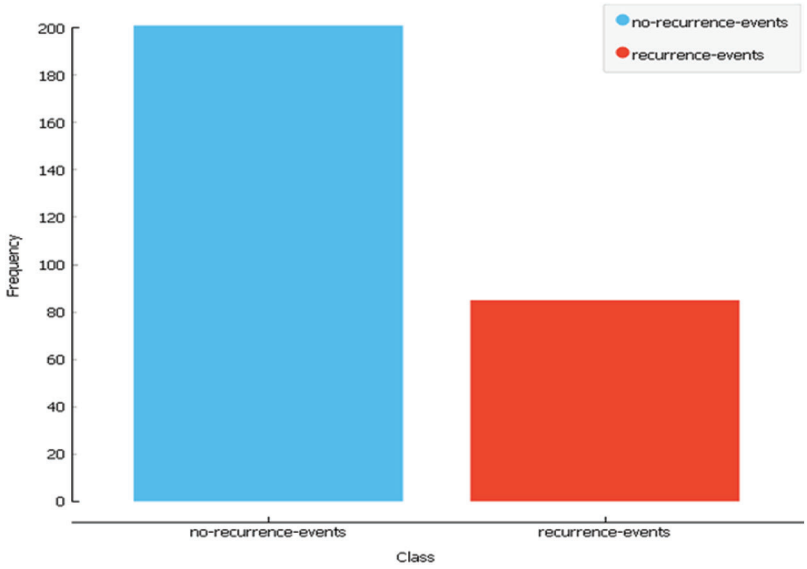

Figure 3: Dataset1 class distribution

4.1.2 Dataset 2

In dataset 2, There are 10 predictors and 116 instances which indicate the presence or absence of breast cancer, all quantitative, and a binary dependent variable. Anthropometric data and parameters that can be obtained by regular blood analysis are the predictors. Prediction models based on these predictors can theoretically be used as a biomarker of breast cancer if they are correct [29]. The following Tab. 2 and Fig. 4 describe the dataset.

Table 2: Dataset 2 attributes

Attributes
Age (years)
BMI (kg/m^2)
Glucose (mg/dL)
Insulin (μU/mL)
HOMA
Leptin (ng/mL)
Adiponectin (μg/mL)
Resistin (ng/mL)
MCP-1 (pg/dL)
Class 1=Healthy controls 2=Patients

4.1.3 Dataset 3

A digitized image of a fine needle aspirate (FNA) of a breast mass measures the characteristics. They define the features of the nucleus of a cell present in the picture. Three-dimensional space is the space defined in: [30]. L. Mangasarian: Robust linear programming Discrimination of Two Linearly Inseparable

Sets. Dataset 3 consists of 31 features and 569 instances. The following Tab. 3 and Fig. 5 describe the dataset. Tab. 4 presents a summary of the three datasets to compare between them.

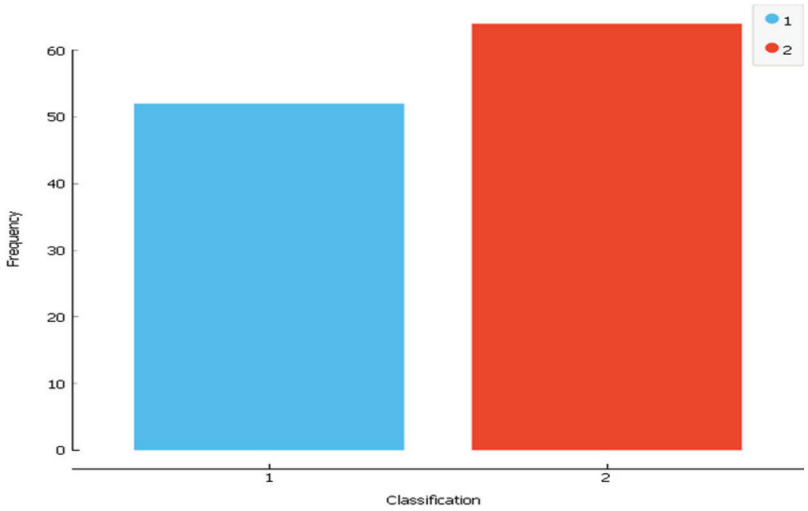

Figure 4: Dataset2 class distribution

Table 3: Dataset 3 features

Features
Radius_worst
Perimeter_worst
Area_worst
Concave points_worst
Concave points_mean
Perimeter_mean
Area_mean
Radius_mean
Concavity_mean
Concavity_worst
Area_se
Perimeter_se
Features
Compactness_mean
Compactness_worst
Radius_se
Texture_worst
Concave points_se
Texture_mean

(Continued)

Table 3 (continued)

Features
Concavity_se
Smoothness_worst
Symmetry_worst
Compactness_se
Smoothness_mean
Symmetry_mean
Fractal_dimension_worst
Fractal_dimension_se
Symmetry_se
Smoothness_se
Texture_se
Fractal_dimension_mean
Diagnose (B or M)

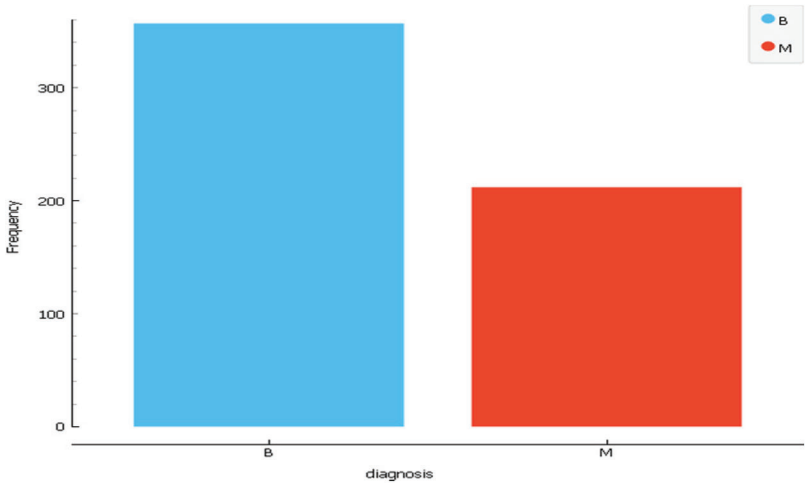

Figure 5: Dataset3 class distribution

Table 4: Presents summary of the datasets

#	Dataset	No. of Instances	Class #1	Class #2
1	Dataset 1	286	85	201
2	Dataset 2	116	64	52
3	Dataset 2	569	212	357

4.2 Classification Model

The proposed model based on three sub-models. Each one is developed by using a dataset and a classification algorithm. As mentioned above, Dataset1, Dataset 2 and Dataset 3 are used to train and test the sub-models. To select the best algorithm in developing the sub-models, intensive experiments have been conducted by using five classification algorithms. The selected algorithms are SVM, Random Forest, Neural Network, Naïve Bayes and Logistic Regression. One of these algorithms will be chosen according to the evaluation process of each sub-model independently. Each sub-model contributes to the final decision to diagnose breast cancer disease. Fig. 6 presents the framework for building the sub-models.

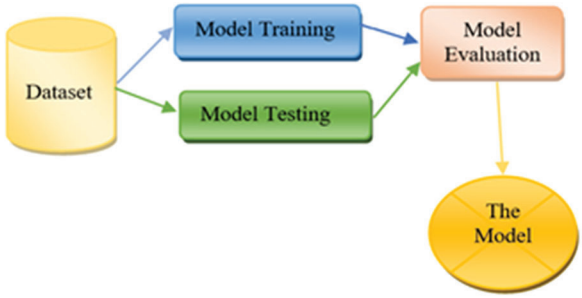

Figure 6: Sub-model framework

4.3 Research Model Implementation

This section presents the proposed model implementation based on the framework in Fig. 2. The model consists of three sub-models. The following sub-sections present the sub-models implementation result.

4.3.1 Sub-Model 1 Implementation

To develop the sub-model 1, Dataset 1 and SVM, Random Forest, Neural Network, Naïve Bayes and, Logistic Regression algorithms were used. Intensive experiments were conducted to select the best algorithms. The following Tab. 5 and Fig. 7 presents the implementation results of Sub-Model 1.

Table 5: Confusion matrix of sub-model 1

Model	AUC	CA	F1	Precision	Recall
SVM	0.666111	0.695804	0.63797	0.644655	0.695804
Random Forest	0.700029	0.674825	0.65436	0.645955	0.674825
Neural Network	0.681592	0.702797	0.696475	0.692275	0.702797
Naive Bayes	0.711853	0.723776	0.721288	0.719255	0.723776
Logistic Regression	0.664501	0.695804	0.668553	0.663799	0.695804

4.3.2 Sub-Model 2 Implementation

To develop the sub-model 2, Dataset 2 and SVM, Random Forest, Neural Network, Naïve Bayes and, Logistic Regression algorithms were used. Intensive experiments were conducted to select the best algorithms. The following Tab. 6 and Fig. 8 presents the implementation results of Sub-Model 2.

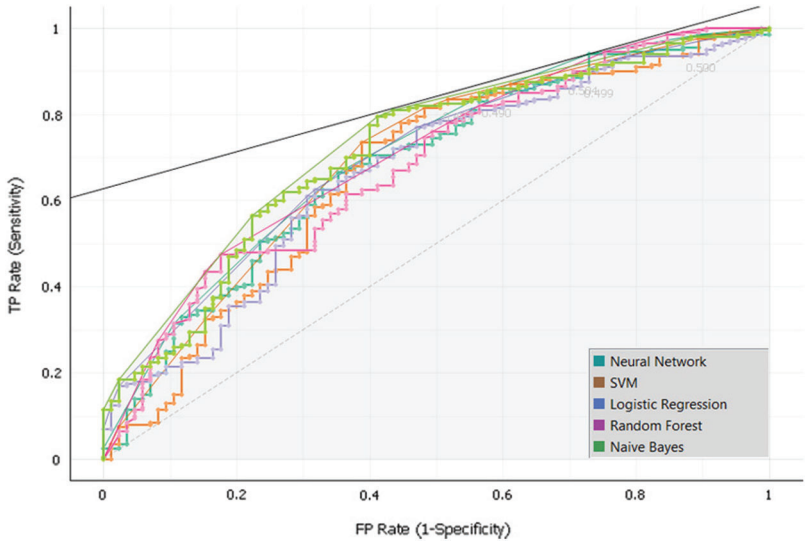

Figure 7: Sub-model 1 ROC

Table 6: Confusion matrix of sub-model 2

Model	AUC	CA	F1	Precision	Recall
SVM	0.82121394	0.74137931	0.74137931	0.74137931	0.74137931
Random Forest	0.74519231	0.62068966	0.62068966	0.62068966	0.62068966
Neural Network	0.8061899	0.70689655	0.70627763	0.7060815	0.70689655
Naive Bayes	0.75510817	0.69827586	0.69906213	0.70493299	0.69827586
Logistic Regression	0.76502404	0.71551724	0.7152389	0.71506735	0.71551724

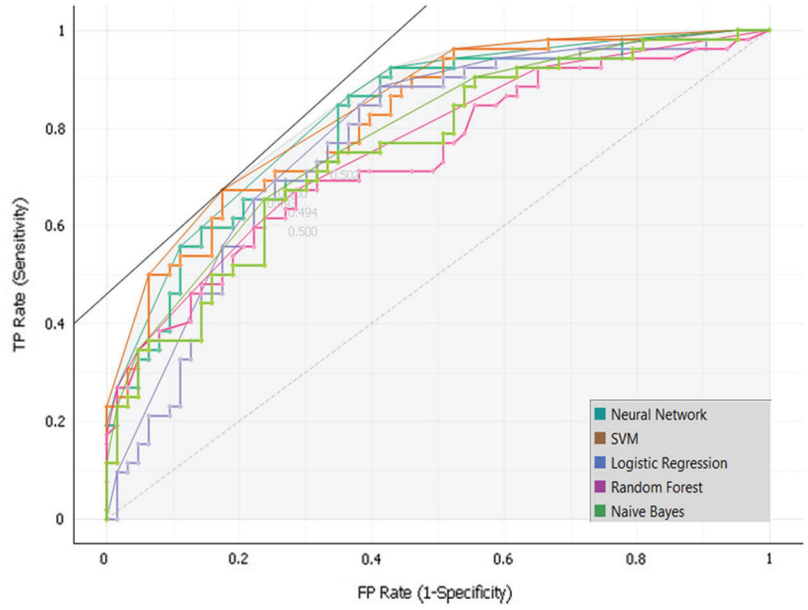

Figure 8: Sub-model 2 ROC

4.3.3 Sub-Model 3 Implementation

To develop the sub-model 3, Dataset 3 and SVM, Random Forest, Neural Network, Naïve Bayes and, Logistic Regression algorithms were used. Intensive experiments were conducted to select the best algorithms. The following Tab. 7 and Fig. 9 presents the implementation results of Sub-Model 3.

Table 7: Confusion matrix of sub-model 3

Model	AUC	CA	F1	Precision	Recall
SVM	0.99467523	0.97363796	0.97362524	0.97361893	0.97363796
Random Forest	0.98923154	0.9543058	0.95416824	0.9542587	0.9543058
Neural Network	0.99151736	0.97188049	0.97185313	0.97185145	0.97188049
Naive Bayes	0.98261191	0.94024605	0.94030239	0.94038327	0.94024605
Logistic Regression	0.47975794	0.60456942	0.52590993	0.53839032	0.60456942

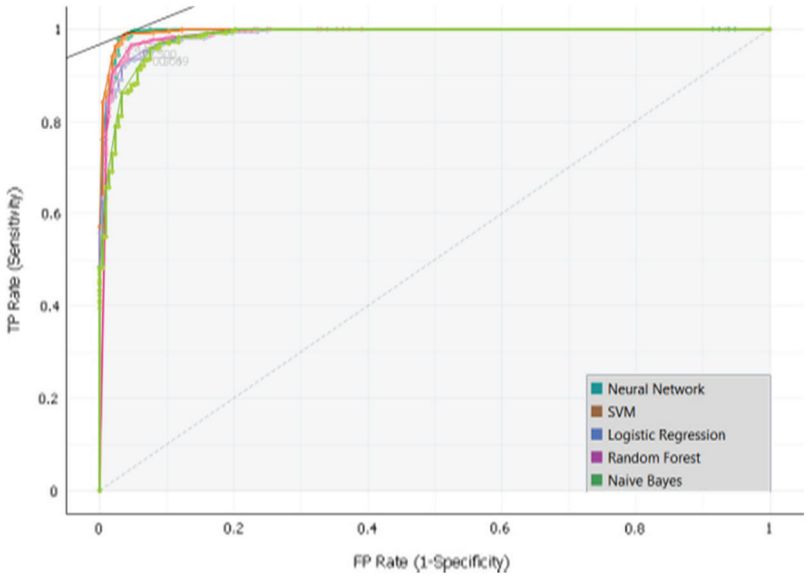

Figure 9: Sub-model 3 ROC

4.4 Results Discussion

As mentioned above, many experiments had been done to determine the research model. In each one, five algorithms were used to create sub-models. Based on sub-model confusion matrixes, one algorithm is select for each sub-model. By this way, the proposed model will be stronger to diagnose breast cancer by different parameters. The following Tab. 8 presents the selected algorithm in each sub-model with obtained results.

Fig. 10 presents the novel classification model for diagnosing breast cancer effectively based on weighting of Heterogeneous Sub-Models.

Table 8: Selected algorithms for sub-models

Sub-Models	Algorithm	AUC	CA	F1	Precision	Recall
1	Naive Bayes	0.711853	0.723776	0.721288	0.719255	0.723776
2	SVM	0.82121394	0.74137931	0.74137931	0.74137931	0.74137931
3	SVM	0.99467523	0.97363796	0.97362524	0.97361893	0.97363796

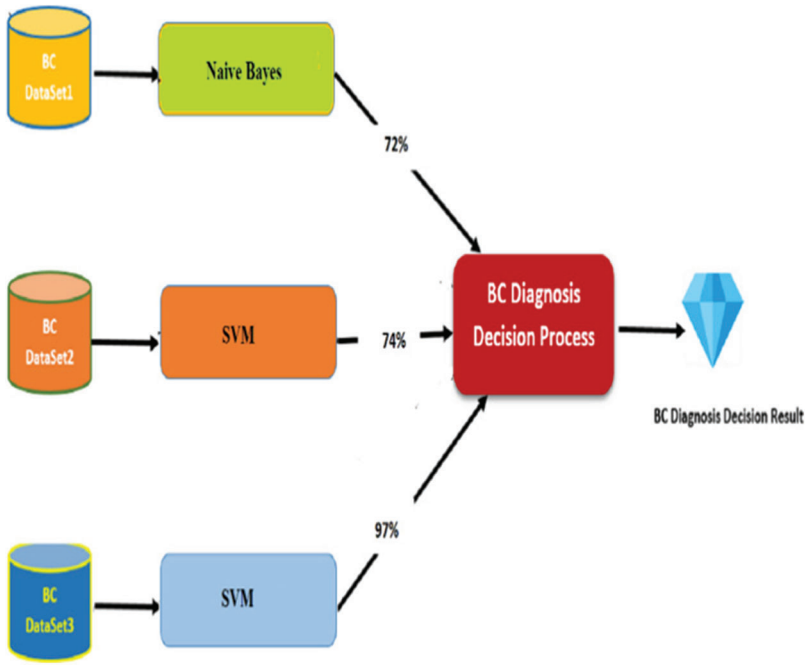

Figure 10: Classification breast cancer model with result

5 Conclusion and Future Work

This paper proposed a novel model for classifying breast cancer patients. The model was developed based on three heterogeneous datasets. The datasets were used to build three sub-models. Each sub-model can diagnose the disease independently. The power of the model comes from the diversity checks of patients and this reduces the risk of wrong diagnosing. Most of the issued models were based on only one dataset which made the diagnosing is not accurate. As we mentioned in related work all proposed models were making a classification based on only one side diagnosing and this disease has high risk, so multiple dimensions of diagnosing should be performed to reach accurate results. The proposed model covers this gap of research. The model has been developed by conducting intensive experiments. Several classification algorithms were used to select the best one in each sub-model. The obtained results have been evaluated and discussed in details.

As a future for this work, the model could be enhanced by adding more features related to lifestyle and social information. In addition, the idea of sub-model could be extended to add more dimensions and that will make the diagnoses accurate.

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through the Small Group Research Project under grant number * (RGP.1/172/42)*

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under grant number *(RGP.1/172/42)*, Received by Majdy M Eltahir. www.kku.edu.sa.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] O. Peart, "Breast intervention and breast cancer treatment options," *Radiologic Technology*, vol. 86, no. 5, pp. 535–558, 2015.

[2] M. K. Derakhshan and M. H. Karbassian, "Psychiatric and psychosocial aspects of breast cancer diagnoses and treatments," *In Cancer Genetics and Psychotherapy*, Springer, Cham, United States, pp. 45–77, 2017.

[3] B. Scholkopf and J. S. Alexander, "Learning with kernels: support vector machines, regularization, optimization, and beyond," *In the Adaptive Computation and Machine Learning Series*, Camberdge Massachusetts, London, England, pp. 15–21, 2018.

[4] M. J. Islam, Q. J. Wu, M. Ahmadi and M. A. Sid-Ahmed, "Investigating the performance of naive-Bayes classifiers and k-nearest neighbor classifiers," *Convergence Information Technology*, vol. 5, no. 2, pp. 133–137, 2010.

[5] M. Etehadtavakol and M. H. Ettefagh, "Evaluation of risk factors in developing breast cancer with expectation maximization algorithm in data mining techniques," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 3, pp. 753–758, 2016.

[6] A. K. Mohanty, M. R. Senapati and S. K. Lenka, "RETRACTED ARTICLE: An improved data mining technique for classification and detection of breast cancer from mammograms," *Neural Computing and Applications*, vol. 22, no. 1, pp. 303–310, 2013.

[7] W. J. Kuo, R. F. Chang, D. R. Chen and C. C. Lee, "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images," *Breast Cancer Research and Treatment*, vol. 66, no. 1, pp. 51–57, 2001.

[8] R. Mousa, Q. Munib and A. Moussa, "Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural," *Expert Systems with Applications*, vol. 28, no. 4, pp. 713–723, 2005.

[9] J. Diz, G. Marreiros and A. Freitas, "Applying data mining techniques to improve breast cancer diagnosis," *Journal of Medical Systems*, vol. 40, no. 9, pp. 1–7, 2016.

[10] V. Santos, N. Datia and M. P. M. Pato, "Classification performance of data mining algorithms applied to breast cancer data," *Computational Vision and Medical Image Processing*, vol. 1234, pp. 307–317, 2013.

[11] R. Shen, Y. Yang and F. Shao, "Intelligent breast cancer prediction model using data mining techniques," in *Proc. of IHMSC*, Hangzhou, China, vol. 1, pp. 384–387, 2014.

[12] N. Mansour, R. Zantout and M. El-Sibai, "Mining breast cancer genetic data," in *Proc. of ICNC*, Shenyang, China, vol. 2, pp. 1047–1051, 2013.

[13] M. R. K. Mookiah, U. R. Acharya and E. Y. K. Ng, "Data mining technique for breast cancer detection in thermograms using hybrid feature extraction strategy," *Quantitative InfraRed Thermography Journal*, vol. 9, no. 2, pp. 151–165, 2012.

[14] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[15] A. S. Sarvestani, A. A. Safavi, N. M. Parandeh and M. Salehi, "Predicting breast cancer survivability using data mining techniques," in *Proc. of ICSTE*, San Juan, PR, USA. vol. 2, pp. V2–227, 2010.

[16] S. Dhahbi, W. Barhoumi, and E. Zagrouba, "Breast cancer diagnosis in digitized mammograms using curvelet moments," *Computers in Biology and Medicine*, vol. 64, no. 1, pp. 79–90, 2015.

[17] S. Turgut, M. Dağtekin and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *Proc. of EBBT*, Istanbul, Turkey, vol. 1, pp. 1–3, 2018.

[18] M. S. Bala and G. R. Lakshmi. "Efficient ensemble classifiers for prediction of breast cancer," *International Journal*, vol. 6, no. 3, pp. 5–9, 2016.

[19] M. A. Abd-Elrazek, A. A. Othman, M. H. Abd Elaziz and M. N. Abd-Elwhab, "Intelligent prediction of breast cancer: A comparative study," *Egyptian Computer Science Journal*, vol. 42, no. 3, pp. 29–43, 2018.

[20] T. Balaraman and V. M. Bhaskaran. "An efficient classifications model for breast cancer prediction based on dimensionality reduction techniques," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, pp. 448–455, 2018.

[21] B. Soni, A. Bora, A. Ghosh and A. Reddy, "RFSVM: A novel classification technique for breast cancer diagnosis," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 12, pp. 3295–3305, 2019.

[22] N. I. Yassin, S. Omran, E. M. El Houby and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 25–45, 2018.

[23] D. Segera, M. Mbuthia and A. Nyete, "Particle swarm optimized hybrid kernel-based multiclass support vector machine for microarray cancer data analysis," *BioMed Research International*, vol. 2019, pp. 1–12, 2019.

[24] R. Senkamalavalli and T. Bhuvaneswari, "Improved classification of breast cancer data using hybrid techniques," *International Journal of Advanced Engineering Research and Science*, vol. 5, no. 5, pp. 77–81, 2017.

[25] S. H. Yesuf, "Breast cancer detection using machine learning techniques," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 5, pp. 27–33, 2019.

[26] S. Gupta, D. Kumar and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian Journal of Computer Science and Engineering (IJCSE)* vol. 2, no. 2, pp. 188–195, 2011.

[27] R. S. PadmaPriya and P. S. Vadivu, "A review on data mining techniques for prediction of breast cancer recurrence," *International Journal of Engineering and Management Research e-ISSN*, pp. 2250–0758, 2019.

[28] M. Zwitter and M. Soklic, "The university medical centre institute of oncology, *In Ljubljana Yugoslavia. UCI Machine Learning Respository*," The University of Massachusetts Amherst, USA, 1988.

[29] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes *et al.,* "Using resistin, glucose, age and BMI to predict the presence of breast cancer." *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018.

[30] K. P. Bennett and L. M. Olvi, "Robust linear programming discrimination of two linearly inseparable sets." *Optimization Methods and Software*, vol. 1, no. 1, pp. 23–34, 1992.