# Feature Selection and the Fusion-based Method for Enhancing the Classification Accuracy of SVM for Breast Cancer Detection

**Ali Ahmed**[†] **and Sharaf J.Malebary**[††]

[†]Department of computer science, faculty of computing and information technology, King Abdulaziz University P. O. Box 344, Rabigh 21911, Saudi Arabia
[†]Faculty of computer science and information technology, Mashreq University, Khartoum North, Sudan
[††]Department of Information Technology, faculty of computing and information technology, King Abdulaziz University P. O. Box 344, Rabigh 21911, Saudi Arabia

**Summary**
Recently, breast cancer has become the second leading cause of death from cancer in women. Although most studies have reported that this form of cancer is preventable and many of the risks can be avoided in its early stages, most of the traditional methods of detecting and diagnosing cancer take place at a very late stage. The classification method is one of the data mining techniques used as a detection method in early stage detection for this type of cancer. Feature selection methods have a positive impact and significant enhancement when used with classification methods. They result in increasing the classification accuracy, since they select the important features of images or any data instances. The objective of this study is to investigate the potential benefit of using the feature selection algorithm as a pre-processing stage for enhancing the classification accuracy of the support vector machine, and to propose a fusion scheme for selecting the best and related features for mammogram images. For this purpose, four feature selection algorithms were chosen, namely mutual information (MI), the statistical dependence measure, the relief-based algorithm and the correlation based algorithm. Extensive experiments have been performed using one of the benchmark datasets, that of the Mammographic Image Analysis Society (MIAS), to test the proposed method on two classes, benign and malignant masses. The results showed that our proposed method at (85 – 15%) data splitting percentage has a classification accuracy of 75% and 93.75% and positive rate of 87.5% and 88.89% for the top seven and top five features, respectively.

***Key words:***
*Machine learning, classification, feature selection, support vector machine, breast cancer detection)*

## 1. Introduction

Feature selection is a significant pre-processing task in data mining and machine learning processes; it has a positive impact in terms of diminishing the information repetitive and high dimensionality of data. Feature selection is defined as a procedure or a process of decreasing features from the data collection that are unimportant regarding the assignment to be performed. Feature selection is significant for several reasons; for example, simplification, performance, computational efficiency and feature interpretability. Feature selection is utilized to fulfill the shared objective and common goal of maximizing the accuracy of the classifier; limiting the related estimation costs; improving precision by lessening unimportant and potentially redundant features; decreasing unpredictability and the related computational expense; and improving the likelihood that a solution will be conceivable and sensible [1, 2].

Feature selection is arranged into two primary classes, filter methods and wrapper methods. In the first group, filter methods select features based on a performance measure regardless of the data modeling algorithm employed. Put simply, after the best features have been discovered, the modeling algorithms can utilize them. Filter methods can rank individual features or assess whole component subsets. The filter utilizes the general attributes of the data itself and work independently from the learning algorithm. More precisely, the filter uses the measurable relationship between a set of features and the target feature. The amount of correlation between features and the target variable determines the significance of the objective or target variable. The characteristics of this method could be summarized thus: they are independent of the classification algorithm, its computational cost is less for a large data set and it executes the task more rapidly compared to the second group [3, 4].

The second group of feature selection methods is known as wrappers. These consider feature subsets by the nature of the presentation and quality of the performance on a modeling algorithm, which is taken as a black box evaluator [5]. Thus, for classification tasks, a wrapper will assess subsets dependent on the classifier performance (e.g. Naïve Bayes or SVM) [6, 7], while for clustering, a wrapper will assess subsets based on the performance of a clustering algorithm (e.g. K-means)[8]. The wrapper evaluates and selects attributes based on precision evaluates by the objective learning algorithm. Utilizing a specific learning calculation, the wrapper essentially looks through the component space by excluding a few features and testing the effect of features oversight on the prediction metrics. The main characteristics of this group

Manuscript received November 5, 2019
Manuscript revised November 20, 2019

of methods are that they depend on the classification algorithm and their computational cost is greater for large data set compared with the first group. However, many of the latest studies have taken advantage of hybrid-based methods for feature selection. This approach, as in [9, 10], uses a combination of both filter and wrapper methods. The following paragraph gives the basic ideas about the most four feature selection algorithms used in this study.

## 2. Related Works

There are many feature selection algorithms found in the literature, and these are widely used to enhance classification accuracy. In the following paragraphs we will give a basic idea of the algorithm used here, which be considered to be as most known feature selection algorithms in this area, along with related studies that use these methods.

### 2.1 Mutual information algorithm

Mutual information (MI) was first introduced by Shannon in 1948 [11]. It is a quantity describing the amount of information two irregular factors or random variables convey about one another. It is symmetric, for example I(X; Y) = I(Y ;X) and ready to recognize non-linear relationships between variables. This last property has made MI a famous model for feature selection since other widely used criteria, such as the correlation coefficient, can only handle linear dependencies. Officially, the MI of a couple of random variables, X and Y, can be defined by means of the probability density function (pdf) of X, Y and the joint variable (X, Y), respectively denoted as $f_X$, $f_Y$ and $f_{X,Y}$

$$I(X;Y) = \iint f_{x,y}(x,y) \log \frac{f_{x,y}(x,y)}{f_x(x) f_y(y)} dxdy \quad (1)$$

If the variables are independent, then $f_{X,Y} = f_X \times f_Y$ and I(X; Y ) = 0.

### 2.2 Statistical dependence measure

In this algorithm, features in specific iteration for an input variable that is most pertinent to the target and least repetitive regarding the effectively chosen variables, are chosen and selected as the best ones. The significance is assessed by the reliance between a variable and the objective, though the repetition is assessed by the normal reliance between the new factor and the effectively chosen variable [12].

### 2.3 Relief-based algorithm

Relief algorithm, inspired by instance-based learning, was developed by Kira and Rendell [13-16]. The authors of this algorithm calculate a proxy statistic for each feature as an individual assessment filtering feature selection method that can be used to predict the 'quality' or 'significance' feature to the target notion (i.e. predicting endpoint value). These feature statistics are referred to as feature weights (W[A] = weight of feature ‘A’), or more casually as feature ‘scores’ that can range from −1 (worst) to +1 (best). The following pseudo-code illustrates the original relief algorithm:

Note: for each training instance a vector of feature values and the class value is found

```
n ← number of instances in training set
a  ← number of attributes or features
Parameter: m  ← number of random training instances out of training set instances used to update W
initialize all feature weights W[A] := 0.0
for i:=1 to m do
Randomly select a ‘target’ instance Ri
  Find a nearest hit ‘H’ and nearest miss ‘M’ (instances)
    for A:= 1 to a do
       W[A]:= W[A]−diff (A,Ri,H)/m+diff (A,Ri,M)/m
     end for
end for
return the vector W of feature scores that estimate the quality of features
```

### 2.4 Correlation based algorithm

The correlation-based selection algorithm, also known as the correlation features selection CFS algorithm, is a straightforward filter algorithm based on a correlation-based heuristic evaluation function that ranks feature subsets [17]. The assessment function's prejudice is towards subsets that contain characteristics that are extremely class-related and uncorrelated. It is important to ignore irrelevant characteristics because they will have low class correlation. Redundant characteristics should be screened out as one or more of the remaining characteristics will strongly correlate them. Acceptance of a function will rely on the extent to which it predicts classes in cases not already predicted by other characteristics in the instance room. The sub-set evaluation function of CFS in the following equation is described below (with slightly altered notation) to facilitate reference:

$$r_{zc} = \frac{k\bar{r}}{\sqrt{k + k(k-1)}} \bar{r}_{ii} \quad (2)$$

where rzc is the correlation between the summed components and the outside variable, k is the number of components, rzi is the average of the correlations between the components and the outside variable, and rii is the average inter-correlation between components.

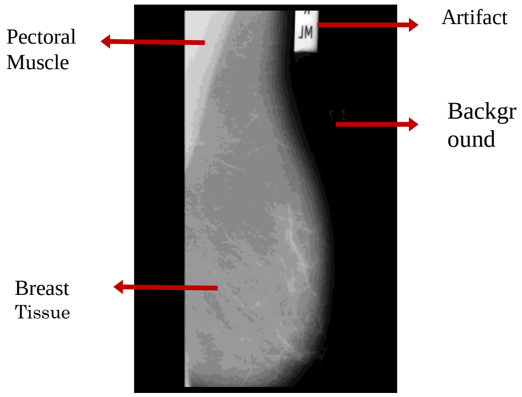

Fig. 2 (a) Sample of Mammogram Image

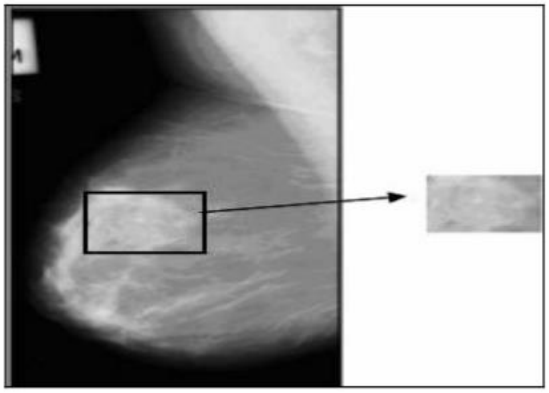

Fig. 2: (b) Sample of Mammogram Image

## 3. Proposed Method

In this study we propose a method for enhancing the classification accuracy of support vector machine SVM. These enhancements are achieved through two steps: first, a group of four common feature selection methods were applied to a dataset to select the most important features. In the second step, the top seven and top five features generated by any of the algorithms are fused or combined together in terms of 2D matrix of 25 or 35 features, respectively. Finally, a group of the most important seven and five features are selected from these fusion features population, then our final classification process will consider only this group of features, and this elimination of features will result in significantly increasing and enhancing the classification accuracy. Our general framework is illustrated in Figure 1.

The dataset used in this study is digital mammogram images collected from the MIAS [18]. It consists of 68 benign images and 51 malignant images. It is a two-class dataset. Figures 2(a) and 2(b) illustrate samples of these images. Before using these images in the further training and testing stages, three important pre-process steps were applied. These phases are: image collection, image cropping based on regions of interest (ROI) and features extraction, as explained in Figure 2.

### 3.1 Image collection and pre-processing

Image processing techniques are applied to images before the feature extraction phase. ROIs are defined as the regions that interest the user based on specific objects defined by the user. We employed the cropping technique to images in order to cut and preserve the interesting parts of the image. Doing so, removed the unwanted parts of the image, usually the surrounding area to the ROI.

### 3.2 Feature extraction and feature selection

In this step, after cropping the (ROI) from [x] to [y] positions and [radius] depend of the MIAS dataset. At this stage we apply the fourteen statistical functions to extract the fourteen feature values from each mammogram image. Samples of these functions are found with their description in Table 1, and it has been used previously by [19].

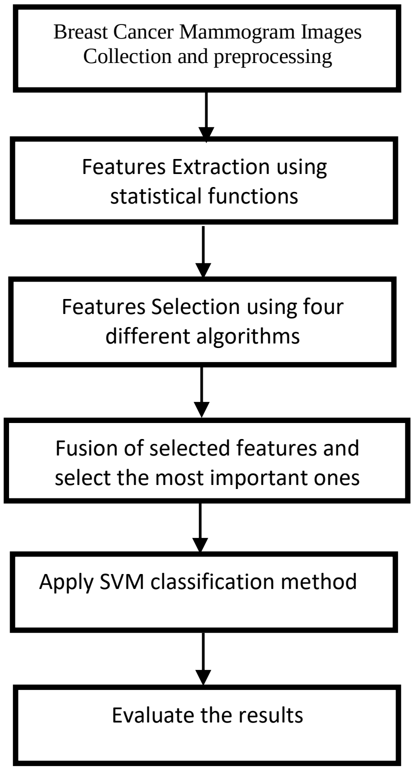
Breast Cancer Mammogram Images Collection and preprocessing
Features Extraction using statistical functions
Features Selection using four different algorithms
Fusion of selected features and select the most important ones
Apply SVM classification method
Evaluate the results

Fig. 1 Proposed framework for enhancing SVM based on Feature Selecting and Fusion Method

Table 1: Sample of feature extraction functions

| **Feature Equation or Formula** | **Description** |
|---|---|
| $Mean = \sum_{i=0}^{L-1} Z_i * p(z_i)$ | “A measure of average intensity” |
| $std = \sum_{i=0}^{L-1} (z_i - m)^2 * p(z_i)$ | “A measure of standard deviation of intensity” |
| $sknewness = \sum_{i=0}^{L-1} (z_i - m)^3 * p(z_i)$ | “Third moment about the mean” |
| $kurtosis = \sum_{i=0}^{L-1} (z_i - m)^4 * p(z_i)$ | “Fourth moment about the mean” |
| $contrast = \sum_{i=0}^{L-1} \sqrt{(z_i - m)^2 * p(z_i)}$ | “Standard deviation of pixel intensities” |
| $smoothness = 1 - \frac{1}{(1+\sigma^2)}$ | “Measures the relative intensity variations in a region” |
| $g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma^2{}_x} + \frac{y^2}{\sigma^2{}_y}\right) + 2\pi j W_x\right]$ | “discrete Gabor wavelet transform” |
| Where “zi is a random variable indicating intensity, p(z) is the histogram of the intensity levels in a region. L is no. of possible intensity levels and $\sigma_\chi$ and $\sigma_\gamma$ are the standard deviation of the Gaussian envelopes along the . x and y” | |

### 3.3 Fusion and selection method

The main idea and hypothesis of this proposed method is that the most important features will give a good representation of the image and will result in good accuracy when this image is chosen for testing stage. Our proposed method is also based on the fact that the group of features selected by a group of features selection algorithms is more representative than features selected by a single algorithm.

Here MI, SD, Relief and CFS feature selection algorithms were applied on the whole dataset of extracted features, then the most important features generated by any of these algorithms were combined into one matrix, and finally the top seven and top five frequent and repeated features only were selected for our final experiments.

### 3.4 Classification based on SVM

The support vector machine (SVM) is a statistical learning theory to analyze data and to recognize patterns [20]. It is a supervised learning method. SVM has some benefits, for instance it can handle continuous and binary attributes. Also the speed of classification and accuracy are good, with few drawbacks i.e., such as SVM takes relatively long time to train a dataset and does not handle discrete attributes well . Here, SVM was applied based on all the feature sets, and then it was applied again on

## 4. Experimental Results and Discussion

In this section we explain our experiments which was accomplished through two main processes and discuss the results. The first part was built for each classifier using the 60, 70, 85 percentages of 119 mammogram images (72, 84 and 95 images) for training purposes. Then, after building the classifier, the remaining 40, 30, 15 percentages (47, 35 and 24 images) of the data set were used in the testing stage. Our experiment was run twice, first based on all the image features and then based only on the selected features and finally the results were presented in the upcoming section. The instance images for training and testing were selected randomly, then all experiments were executed ten times and the average was calculated. To test the performance of the proposed method, recall and classification accuracy quantitative measures have been used; both of these can be calculated by using the following two equations, while the results of our experiments are illustrated in Table 2.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3}$$

$$recall = \frac{TP}{(TP + FP)} \tag{4}$$

Where TP is the True Positive, FP is the False Positive, FN is the False Negative and TN is the True Negative.

Table 2 : Classification accuracy for individual FS algorithms and proposed method

| | **TOP 7 Features** | | | | | |
|---|---|---|---|---|---|---|
| | 60-40% | | **70-30%** | | **85-15%** | |
| | TPRate | ACC | **TPRate** | **ACC** | **TPRate** | **ACC** |
| **MI** | 0.5556 | 0.5952 | 0.5000 | 0.5313 | 0.7143 | 0.3750 |
| **SD** | 0.5000 | 0.5714 | 0.4286 | 0.6250 | 0.7143 | 0.7500 |
| **Relieff** | 0.5000 | 0.6190 | 0.5000 | 0.5938 | 0.7143 | 0.6875 |
| **CFC** | 0.3889 | 0.6667 | 0.3571 | 0.6250 | 0.5714 | 0.6875 |
| **Fusion and selection** | 0.5556 | 0.6905 | 0.5714 | 0.6875 | 0.8750 | 0.7500 |
| | TOP 5 Features | | | | | |
| | 60-40% | | 70-30% | | 85-15% | |
| | **TPRate** | **ACC** | **TPRate** | **ACC** | **TPRate** | **ACC** |
| **MI** | 0.6667 | 0.5714 | 0.5714 | 0.5938 | 0.8571 | 0.5625 |
| **SD** | 0.6000 | 0.6200 | 0.6350 | 0.6300 | 0.5714 | 0.7500 |
| **Relieff** | 0.6667 | 0.6190 | 0.7143 | 0.6250 | 0.8571 | 0.8125 |
| **CFC** | 0.4444 | 0.6667 | 0.5714 | 0.6250 | 0.7777 | 0.6250 |
| **Fusion and selection** | 0.7222 | 0.7143 | 0.7857 | 0.7000 | 0.8888 | 0.9375 |

From the results achieved in this study, we can observe that the SVM classification method has good performance in terms of accuracy and recall when it is based only on selected features chosen by our proposed method. From the results shown in Table 2, we can observe that the group of features generated by this fusion method is better than any group of filtering features generated by any of the four feature selection algorithms separately.

From the results shown in Table 2, we can observe that the group of features generated by this fusion method is better than any group of filtering features generated by any of the four feature selection algorithms separately.

Another two measures, precision or positive predictive value (PPV) and miss rate or false negative rate (FNR), for the top five features when the data is split into 85% and 15%, are computed and presented in Figure 3. Again, it is clearly noted that important features filtered by our proposed method have high precision and low false rate compared with the features generated by the four algorithms. These two measures are computed using the following equations:

$$PPV = \frac{TP}{TP + FP} \qquad (5)$$

$$FNR = \frac{FN}{FN + TP} \qquad (6)$$

## 5. Conclusion

Many researchers have concluded that the reduction and elimination of redundant features results in good classification accuracy, but the way of selecting these important features differs from one algorithm to another. The fusion-based method for feature selection proposed here selects and generates the most important features since it has higher classification when used with the SVM classifier.

### Acknowledgment

This work was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (D: 1440-186-830). The authors, therefore, gratefully acknowledge the DSR for their technical and financial support.

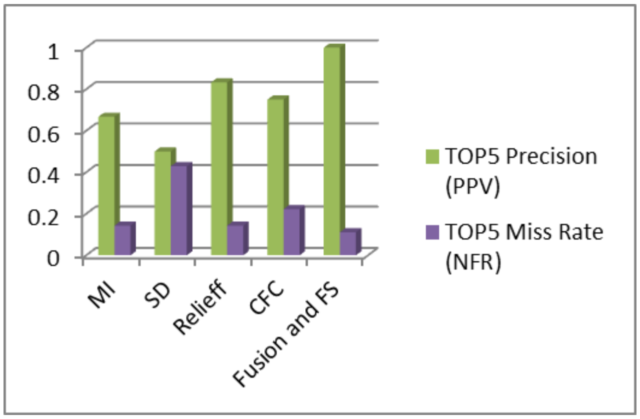

Fig 3 Precision and false negative

## References

[1] Cai, J., J. Luo, S. Wang and S. Yang, Feature selection in machine learning: A new perspective. Neurocomputing, 2018. 300: p. 70-79.

[2] Sheikhpour, R., M. A. Sarram, S. Gharaghani and M. A. Z. Chahooki, A survey on semi-supervised feature selection methods. Pattern Recognition, 2017. 64: p. 141-158.

[3] Novaković, J., Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 2016. 21(1).

[4] Jović, A., K. Brkić, and N. Bogunović. A review of feature selection methods with applications. in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2015. IEEE.

[5] Mafarja, M. and S. Mirjalili, Whale optimization approaches for wrapper feature selection. Applied Soft Computing, 2018. 62: p. 441-453.

[6] Bradley, P.S. and O.L. Mangasarian. Feature selection via concave minimization and support vector machines. in ICML. 1998.

[7] Maldonado, S., R. Weber, and F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. Information Sciences, 2014. 286: p. 228-246.

[8] Kim, Y., W.N. Street, and F. Menczer, Evolutionary model selection in unsupervised learning. Intelligent data analysis, 2002. 6(6): p. 531-556.

[9] Yan, C., J. Liang, M. Zhao, X. Zhang, T. Zhang and H. Li, A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. Analytica chimica acta, 2019. 1080: p. 35-42.

[10] Wang, Y. and L. Feng, Hybrid feature selection using component co-occurrence based feature relevance measurement. Expert Systems with Applications, 2018. 102: p. 83-99.

[11] Shannon, C.E., A mathematical theory of communication. Bell system technical journal, 1948. 27(3): p. 379-423.

[12] Bolón-Canedo, V., S. Seth, N. Sánchez-Marono, A. Alonso-Betanzos and J. C. Prı́ncipe, Statistical dependence measure for feature selection in microarray datasets. in ESANN. 2011.

[13] Kira, K. and L.A. Rendell, A practical approach to feature selection, in Machine Learning Proceedings 1992. 1992, Elsevier. p. 249-256.

[14] Kira, K. and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. in Aaai. 1992.

[15] Aha, D.W., D. Kibler, and M.K. Albert, Instance-based learning algorithms. Machine learning, 1991. 6(1): p. 37-66.

[16] Callan, J.P., T. Fawcett, and E.L. Rissland. CABOT: An Adaptive Approach to Case-Based Search. in IJCAI. 1991.

[17] Hall, M.A., Correlation-based feature selection for machine learning. 1999.

[18] http://archive.ics.uci.edu/ml/datasets/mammographic+mass: access on October 2019

[19] Aarthi, R., Divya, K., Komala, N., Kavitha, S.: Application of Feature Extraction and Clustering in Mammogram Classification using Support Vector Machine, In: International Conference on Advanced Computing (ICoAC), pp. 62–67, IEEE (2011).

[20] Arning, A., R. Agrawal, and P. Raghavan. A Linear Method for Deviation Detection in Large Databases. in KDD. 1996.

**Dr. Ali Ahmed** is an Associate Professor at King Abdulaziz University – Rabigh.. He received his B.Sc. from Karary University (Sudan) in computer engineering in 2001 and his M.Sc. degree in computer science, from Khartoum University (Sudan) in 2003. He received his PhD degree in computer science and his Post-doctoral from UTM University (Malaysia) in 2013 and 2014 respectively. After working as an assistant professor in Sudan (since 2014), now he is an associate professor at FCITR in king abdulaziz university in KSA. His research interest includes machine learning, information retrieval and medical image retrieval

**Dr. Sharaf J. Malebary** is an Assistant Professor at King Abdulaziz University – Rabigh. He received his PhD in Computer Science and Engineering from University of South Carolina, USA. His research interest is Autonomous Systems, Wireless Communications, Artificial Intelligence, and Machine Learning.

Dr. Malebary is very enthusiastic individual and has shown and proven strong leadership potentials. As a result, he serves as Head of Information Technology Department. Moreover, He was nominated to serve as Vice-Dean for Graduate Studies and Scientific Research at the Faculty of Computing and Information Technology at King Abdulaziz University, Rabigh Branch.