Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets

Article *in* Journal of the Chinese Institute of Engineers · October 2019

DOI: 10.1080/02533839.2019.1676658

CITATIONS
45

READS
1,263

4 authors:

Zohaib Mushtaq
University of Sargodha
37 PUBLICATIONS **533** CITATIONS
SEE PROFILE

Akbari Yaqub
National University of Computer and Emerging Sciences
4 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Shaima Sani
5 PUBLICATIONS **74** CITATIONS
SEE PROFILE

Adnan Khalid
University of Management and technology lahore. Sialkot Campus
8 PUBLICATIONS **107** CITATIONS
SEE PROFILE

Journal of the Chinese Institute of Engineers

ISSN: 0253-3839 (Print) 2158-7299 (Online) Journal homepage: https://www.tandfonline.com/loi/tcie20

Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets

Zohaib Mushtaq, Akbari Yaqub, Shaima Sani & Adnan Khalid

To cite this article: Zohaib Mushtaq, Akbari Yaqub, Shaima Sani & Adnan Khalid (2019): Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets, Journal of the Chinese Institute of Engineers, DOI: 10.1080/02533839.2019.1676658

To link to this article: https://doi.org/10.1080/02533839.2019.1676658

Published online: 22 Oct 2019.

JOURNAL OF THE CHINESE INSTITUTE OF ENGINEERS
https://doi.org/10.1080/02533839.2019.1676658

Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets

Zohaib Mushtaq[a], Akbari Yaqub[b], Shaima Sani[c] and Adnan Khalid[d]

[a]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan; [b]Department of Electrical Engineering, FAST University of Computer and Emerging Science, Lahore, Pakistan; [c]School of Informatics Science, UMT, Lahore, Pakistan; [d]Department of Electrical Engineering, UMT, Sialkot, Pakistan

ABSTRACT

Using machine learning algorithms for early prediction of the signs and symptoms of breast cancer is in demand nowadays. One of these algorithms is the *K*-nearest neighbor (KNN), which uses a technique for measuring the distance among data. The performance of KNN depends on the number of neighboring elements known as the *K* value. This study involves the exploration of KNN performance by using various distance functions and *K* values to find an effective KNN. Wisconsin breast cancer (WBC) and Wisconsin diagnostic breast cancer (WDBC) datasets from the UC Irvine machine learning repository were used as our main data sources. Experiments with each dataset were composed of three iterations. The first iteration of the experiment was without feature selection. The second one was the *L*1-norm based selection from the model, which used the linear support vector classifier feature selection, and the third iteration was with Chi-square-based feature selection. Numerous evaluation metrics like accuracy, receiver operating characteristic (ROC) curve with the area under curve (AUC) and sensitivity, etc., were used for the assessment of the implemented techniques. The results indicated that the technique involving the Chi-square-based feature selection achieved the highest accuracy with the Canberra or Manhattan distance functions for both datasets. The optimal *K* values for these distance functions ranged from 1 to 9. This study indicated that with the appropriate selection of the *K* value and a distance function in KNN, the Chi-square-based feature selection for the WBC datasets gives the highest accuracy rate as compared with the existing models.

Abbreviations: KNN: K-nearest neighbor; Chi2: Chi-square; WBC: Wisconsin breast cancer

ARTICLE HISTORY
Received 14 March 2019
Accepted 17 September 2019

KEYWORDS
K-nearest neighbor; Wisconsin breast cancer; feature selection; Chi-square

1. Introduction

One of the most common types of cancer is breast cancer. Both men and women can be affected by breast cancer, but it is observed that the women are the main victims of this disease. Out of 266120 estimated new cases, 40920 deaths were predicted in the United States during 2018 (Institute, National Cancer for surveillence, epidemology and End result program 2018). However, early treatment and diagnosis can save many lives. In this regard, many machine learning techniques have already been implemented to diagnose breast cancer with good accuracy rates as reported by Yue et al., in 2018;Salama, Abdelhalim, and Zeid in 2012. The accomplishment of such techniques can be tested with various performance assessment metrics.

The *K* nearest neighbor (KNN) is considered as one of the most famous machine learning algorithms for diagnosing breast cancer. Hence, most of the current techniques are based on the idea of finding an effective KNN for diagnosing breast cancer. The Wisconsin breast cancer (WBC) and the Wisconsin diagnostic breast cancer (WDBC) datasets from the UC Irvine machine learning repository supported by the University of California are the major source for obtaining data. In order to obtain an effective breast cancer classification, a number of approaches, such as Artificial Neural Network (ANN) with a Feed Forward Back Propagation Neural Network (FFBPNN), have previously been applied to the WBC dataset, and the highest achieved accuracy was found to be 98% (Abdel-Ilah and Šahinbegovi 2017). A novel generalized flow using the cloud-based Microsoft Azure Machine Learning Studio (MAMLS) platform was implemented for various datasets, which includes the WBC dataset. In addition, logistic regression, neural networks, support vector machines (SVM), and decision forest models have also been used on the WBC dataset. The proposed generalized flow approach in (Bihis and Roychowdhury 2015) achieved accuracy ranges from 78–97.5%. A computer aided diagnosis (CAD) scheme for breast cancer diagnoses has also been developed. In order to achieve a better result, the deep belief neural network (DBNN) with the Levenberg Marquardt (LM) learning function was used. But with deep learning, there is a risk of overfitting, and it is very complex to train even a simple network as training a model consumes a great deal of time (Abdel-Zaher and Eldeib 2016). The Chi-square-based important features selection was employed in (Juneja and Rana 2018) to classify cancerous cells from the WBC and the WDBC datasets. These features were utilized by a few machine learning algorithms where naïve Bayes, decision tree, random forest, random tree, and weighted decision tree were implemented for better detection of tumorous cells. An ANN was implemented by using

CONTACT Zohaib Mushtaq engr_zohaibmushtaq@yahoo.com Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

various techniques such as the back propagation algorithm (BPA), radial basis function (RBF), and multi-layer perceptron (MLP), and the best-achieved accuracy was 99% (Osmanović et al. 2019). The rough set-based support vector machine (RS_SVM) algorithm was proposed in (Chen et al. 2011). The reported accuracy was 96.87%.

Various classifiers have been proposed for the WDBC (You and Rumbe 2010) where KNN shows the best result compared to SVM and naïve Bayes. In another study, ANN, SVM, and the semi-supervised models were used for the breast cancer dataset. It is found that the semi-supervised learning model performs better compared with other models reported (Park et al. 2013). The comparative study and analysis of machine learning algorithms were also discussed previously (Asri et al. 2016). SVM, naïve Bayes, *C*4.5 based decision tree, and KNN were applied on the WBC dataset in that study. The SVM classifier achieved 97.13% accuracy for the WBC dataset. Rathi (2016) stated that the minimum redundancy maximum relevancy-based (MRMR) feature selection could be applied to the breast cancer dataset. Various error metrics like accuracy, sensitivity, and specificity were used in their study. The maximum accuracy reached 99% with the use of SVM. In the breast cancer cellular images dataset (Tripathy, Mahanta, and Paul 2014), four different algorithms were applied for accurate breast cancer classification, which includes KNN, SVM, ANN, and the least-square support vector machine (LS-SVM). From the results, it is found that the radial basis function and the linear kernel-based LS-SVM exhibited the highest achieved accuracy of 95.34%. An ensemble-based feature selection technique entitled RMean was proposed. This model provided improvements on the results of the SVM and the FFBPNN with AUC scores of 0.777 and 0.944, respectively (Pérez, Frias, and Silva 2015). Various feature selecting techniques have also been implemented on Weka. The experiment showed that the wrapper technique gave the best accuracy for LS-SVM compared to others (Hall and Holmes 2003). Four different distance functions were used to check their effects on the accuracy of KNN. The datasets used for their study were categorical, numerical, and mixed data types. The results showed that the Chi-square performed the best among them (Hu et al. 2016).

The aim of this study is to classify malignant and benign tumors from the WBC and WDBC datasets. Hence, eight different distance functions were considered for KNN, employed on these datasets. These distance functions are Canberra, Manhattan, Hamming, Chebyshev, Euclidean, Correlation, Minkowski and Cosine, respectively, with *K* values ranging from 1 to 59. The Wisconsin breast cancer (WBC) and the Wisconsin diagnostic breast cancer (WDBC) datasets from the UC Irvine machine learning repository supported by the University of California were used as the primary datasets of this study. In order to obtain an effective KNN, feature selection approaches have also been considered. Furthermore, the *L*1 based linear support vector classifier feature selection and the Chi-square-based feature selection techniques have been applied to these datasets. The findings of our study indicated that the performance of KNN can significantly be improved after the application of feature selection techniques. In addition, the Chi-square proved to be a good feature selection technique for the WBC and the WDBC datasets. The Manhattan and the Canberra distance functions illustrated exceptional performance on these datasets. The best K values for the aforementioned distance functions obtained in this study were 1, 8, and 7.

The methodology reported in this manuscript is one of the best in terms of accuracy, training and prediction time of models for the WBC and WDBC datasets. This technique is lightweight, quick, effective and efficient. It does not involve any complex algorithms with hybrid approaches or time-consuming algorithm fusion techniques. Hence, this approach provides a novel idea for utilizing the maximum possible distance functions with suitable *K* values ranging from 1 to 59 for KNN. The optimal *K* value and distance function were organized and accompanied by preprocessing and suitable feature selection techniques. Later, these parameters were trained and tested for both the WBC and the WDBC datasets.

Feature selection techniques have also been applied to obtain better results. Each experiment was repeated several times in our tests. These results were compared with the results of other existing models. The organization of this study is as follows: Section 2 gives the description of the datasets. Section 3 states the proposed methodology. Section 4 shows the experimental results and discussion. Section 5 is the comparison of the results with other existing models, and Section 6, is the conclusion.

2. Datasets description

This study uses two different datasets from the UCI repository (Wolberg, Street, and Mangasarian 2019). They are the WBC and the WDBC datasets. WBC consists of 699 instances with 10 attributes columns and one class column. There are 16 missing values in the bare nuclei column denoted by '?' as shown in Table 1. The class column has two classes, being four (malignant), which is 65.5% of the total, and two (benign) contributing 34.5%.

The second dataset is the WDBC. These features are better understood through the fine needle aspiration technique of a mass taken from the breast, and Dr. William is the creator of this dataset. The total consists of 569 instances with 32 features as shown in Table 2. The diagnosis column consists of class 'M' as malignant and class 'B' as benign. The class distribution in WDBC are 357 instances as benign and 212 instances as malignant. Among these features, the simple ID number is not used, and the remaining 10 entries are real values. The mean, the

Table 1. WBC dataset description.

Attribute #	Wisconsin breast cancer (WBC) features	Range of values
1	Sample code number id number	Special Id Number
2	Clump thickness	1-10
3	Uniformity of cell size	1-10
4	Uniformity of cell shape	1-10
5	Marginal adhesion	1-10
6	Single epithelial cell size	1-10
7	Bare nuclei	1-10 (missing values '?')
8	Bland chromatin	1-10
9	Normal nucleoli	1-10
10	Mitoses	1-10
11	Class	2 'Benign' and 4 'Malignant'

Table 2. WDBC dataset description.

Features #	Features names	Features domain	Features measurement range		
			Mean	Standard error	Maximum
1	ID	Real	–	–	–
2	Diagnoses	Real	–	–	–
3	Radius	Real	6.98–28.11	0.112–2.873	7.93–36.04
4	Texture	Real	9.71–39.28	0.36–4.89	12.02–49.54
5	Perimeter	Real	43.79–188.5	0.76–21.98	50.41–251.20
6	Area	Real	143.50–2501	6.80–542.20	185.20–4254
7	Smoothness	Real	0.053–0.163	0.002–0.031	0.071–0.223
8	Compactness	Real	0.019–0.345	0.002–0.135	0.027–1.058
9	Concavity	Real	0.000–0.427	0.000–0.396	0.000–1.252
10	Concave points	Real	0.000–0.201	0.000–0.053	0.000–0.291
11	Symmetry	Real	0.106–0.304	0.008–0.079	0.157–0.664
12	Fractal dimension	Real	0.050–0.097	0.001–0.030	0.055–0.208

standard error, and the maximum value for these three sub-columns for the 10 features are as given in this dataset. Unlike WBC, WDBC does not have any missing values.

3. Methodology

KNN is the most widely used algorithm for classification tasks. The performance of KNN depends upon the *K* value and the distance function used in the algorithm. The whole process of this study involves a few steps, as outlined in Figure 1. The first step is preprocessing, which enables the missing values by using the mean value of that feature. Next is the feature scaling, as the min-max based feature scaling is implemented. *L*1 based selection from the model and the Chi-square-based feature selection are both considered to see which one can provide better performance. Different performance evaluation metrics have been used for the analysis of the KNN classifier with *K* values ranging from 1 to 59 and various distance function parameters. The experiment repeats for each distance function on each dataset. Different *K* values and all the distance functions discussed in section 3.3 were employed to obtain the best accuracy. Various performance evaluation metrics like the ROC curve with the AUC, Matthew correlation coefficient (MCC), sensitivity, selectivity, f1-score, error-rate, miss-rate, and fall-out have been used. The whole process repeats with the use of *L*1 based selection from the model feature selection. Support vector classifier (SVC) was used to select suitable features. Another technique is the Chi-square-based feature selection. The same procedure repeats for both the WBC and the WDBC datasets.

3.1. Data preprocessing

As aforementioned, the first step is preprocessing. It is certain that cleaning data and dealing with the missing values can

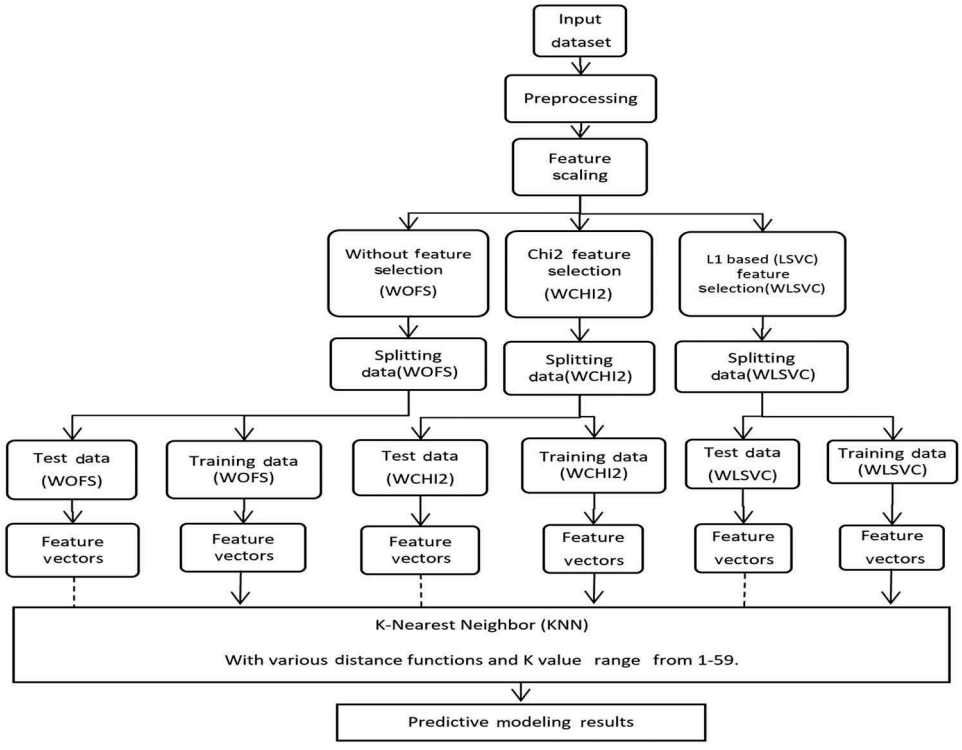

Figure 1. General block diagram of the proposed study.

improve the classification accuracy. This preprocessing step first involves removing useless columns or entries. In the datasets used in this study, the first columns of these datasets are sample ID numbers, which are useless terms for our predictive models. In this step, these columns are removed from the datasets. Processing missing values is also an important task for data mining. Skipping these missing values is not a good solution, as valuable information can also be lost. An effective solution for this type of problem is to replace the missing values with the mean of the entries in that column. These small steps have the ability to boost the performance of the classifiers.

Another data preprocessing step is feature scaling. Min-max based feature scaling was implemented to find the minimum and maximum values of all entries for the considered feature. Then those values were normalized into a specific interval. Since the considered feature values were all positive, the normalized interval was [0, 1].

3.2. Feature selection

Feature selection is a very important technique to improve the performance of any classifier. With the help of the feature selection, the average training and predicting time can be reduced. In this process, only the most important features have been selected from the actual datasets. Later, these selected features were considered for training and testing. Such techniques have a large impact on the classification results.

3.2.1. L1 based selection from model feature selection

The first feature selection considered in our study is the *L*1 based feature selection. It gives zero weights to insignificant features and non-zero weights to important features. This approach can be used with classifiers to achieve dimensionality reduction for given datasets. Sparse estimators for this task play a vital role in satisfactory performance. Logistic regression and SVM are usually used for the classification task. In this study, the linear support vector classifier (LSVC) was used, and a C parameter was adopted to control the sparsity. It can be observed that the larger the value of C, the more features will be selected, and the smaller the value of C, the fewer features will be selected.

This *L*1 based selection from the model feature selection technique has been applied in this article with the help of the scikit-learn library in Python. It is very easy to use and is written in a high-level language with a remarkable response time of various algorithms and technique compared with other existing libraries like mlpy, pybrain, shogun, etc. This package involves a lot of supervised and unsupervised algorithms with different feature selection techniques and methodologies. This library is also well -associated with other famous Python libraries like pandas, NumPy, matplotlib, etc. Non-programmers and beginners can easily take advantage of this library to utilize the various machine learning algorithms, Pedregosa, Weiss, and Brucher (2011).

3.2.2. Chi-square-based feature selection

Feature selection by using the Chi-square technique involves the Chi-square test. This test is based upon a statistical test, which is used to better understand the dependency of one variable on another variable. In this case, suppose we have a target variable, which is also known as a class variable, and other variables are feature variables involved in our data. In this Chi-square, the relationship between all the feature variables and the target variable are observed. If there are some feature variables which do not have a dependency relationship with the target variable, then these features will be eliminated from the feature selection list. If a few feature variables and the target variable depend more upon each other than those feature variables; they are very important and will be selected.

X_n = Total number of instances

X_p = Number of positive instances which contain feature X

X_{neg} = Number of negative instances which contain feature X

$X_{p\_}$ = Number of positive instances which do not contain feature X

$X_{neg\_}$ = Number of negative instances which do not contain feature X

$X_{pn} = X_p + X_{neg}$ = Total number of instances which contain feature X

$X_{nn} = X_{p\_} + X_{neg\_} = X_n - X_{pn}$ = Instances which do not contain feature X

we can have the following calculation.

$$X_p' = \frac{(X_p + X_{p\_})(X_p + X_{neg})}{X_n}, \tag{1}$$

$$X^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}, \tag{2}$$

X^2 = Chi-squared test, O = Observed value, E = Expected value

$$X^2 = \frac{X_n\left((X_p * X_{neg\_}) - (X_{neg} * X_{p\_})\right)^2}{(X_p + X_{p\_})(X_{neg} + X_{neg\_})(X_p + X_{neg})(X_{p\_} + X_{neg\_})}, \tag{3}$$

Table 3 describes the selection of features after preprocessing and scaling for each dataset. The original dataset for the WBC involves 11 columns and includes 32 columns for the WDBC. The ID number and class column in each data set were eliminated. In the case of the WBC, two more features were dropped by using the Chi-square feature selection as they are independent of our class variables. *L*1 based feature selection dropped three features which were less related to the target variable. For the WDBC dataset, 10 unimportant features were discarded from the features list with the Chi-square (WCHI2) feature selection technique. In the same scenario, 16 unimportant features are not included by using the linear support vector classifier (WLSVC) feature selection. Feature selection technique plays a vital role in the overall efficiency of the model by reducing the training time.

3.3. K-nearest neighbor (KNN)

KNN is a simple classifier whose performance depends upon the similarity measures (distance functions) used and the

Table 3. Numbers of features selected in techniques.

Dataset name	Without feature selection (WOFS)	Chi2 feature selection (WCHI2)	LSVC feature selection (WLSVC)
(WBC)	9	7	6
(WDBC)	30	20	14

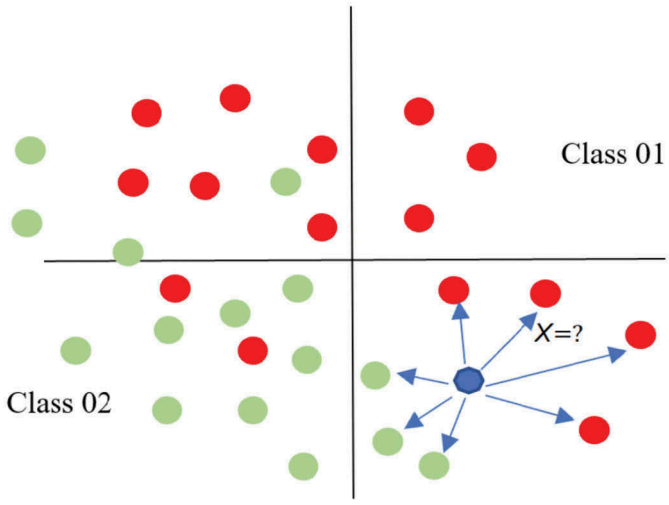

Figure 2. K nearest neighbor working principle.

K value selected. KNN is normally used for classification problems to predict discrete values. It can also be useful for predicting continuous values which belong to regression tasks (Li et al. 2012). Figure 2 shows the working principle of KNN. The red circles belong to class01, and green circles are associated with class02. The blue circle is the point which is to be predicted by the algorithm, either it is related to class01 or class02. Two factors are very crucial in KNN. One is the distance function, and the other is the *K* value, which demonstrates the number of neighboring dots or circles close to our target blue point. Another important term is the distance function, which is an approach of finding the closest distance between the targeted blue circle and other neighboring instances. In Figure 2, symbol '*X*' denotes a distance function, and the blue arrows target the neighboring dots of each class close to the blue circle. If our target blue dot is close to a larger number of elements or circles from class01, then KNN will predict the blue circle as a class01 element; otherwise, the algorithm will predict the blue point as a class02 element.

In KNN, there is a measurement about distance. Various distance functions are employed in this study. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be feature vectors. $\boldsymbol{a} = (a_1, a_{2,\ldots}. a_n)$ and $\boldsymbol{b} = (b_1, b_{2,\ldots}\ldots b_n)$. Those considered distance functions are discussed as follows.

$$Minkowski_{d(a,b)} = \left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{\frac{1}{p}} \quad where\ p = 1, 2, \ldots\ldots\ldots\infty, \tag{4}$$

$$Euclidean_{d(a,b)} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots\cdots (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}, \tag{5}$$

$$Manhattan_{d(a,b)} = \sum_{i=1}^{n} |a_i - b_i|, \tag{6}$$

$$Hamming_{d(a,b)} = def \left\{ \begin{array}{l} 0\ if\ a = b \\ 1\ otherwise \end{array} \right\}, \tag{7}$$

$$Cosine_{d(a,b)} = \sum_{i=1}^{n} (a_i)(bi) / \sqrt{\sum_{i=1}^{n} (a_i)^2} \sqrt{\sum_{i=1}^{n} (b_i)^2}, \tag{8}$$

$$Canberra_{d(a,b)} = \sum_{i=1}^{n} (|a_i - b_i| / |a_i| + |b_i|), \tag{9}$$

$$Correlation_{d(a,b)} = cov(a, b) / (\sigma a)(\sigma b) = 1 - c_{ai,bi}, \tag{10}$$

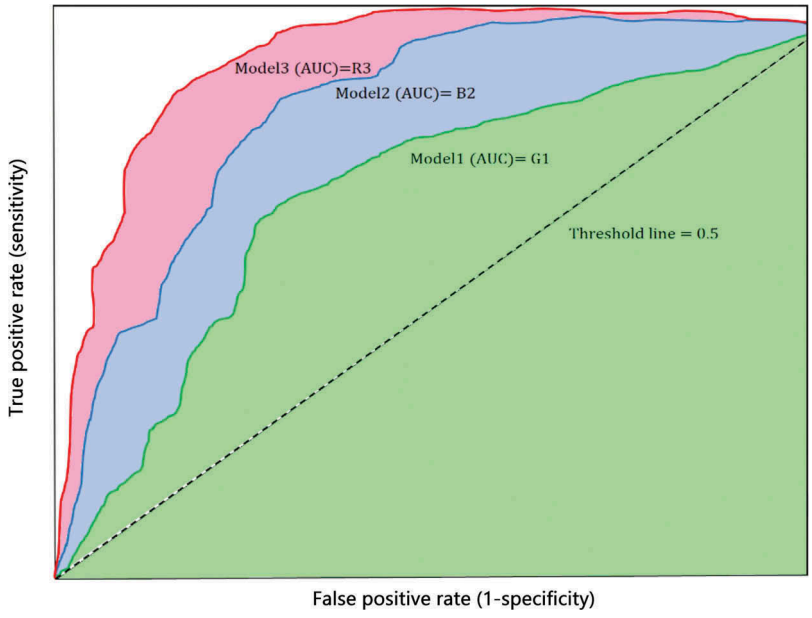

Figure 3. Explanation of receiver operating characteristic (ROC) and area under curve (AUC) .

Table 4. Training and predicting time in seconds for the Wisconsin breast cancer (WBC) dataset.

KNN model distances	Training time WOFS(s)	Predicting time WOFS(s)	Training time WLSVC(s)	Predicting time WLSVC(s)	Training time WCHI2(s)	Predicting time WCHI2(s)
Euclidean	1.586	0.827	1.09	0.755	0.962	0.741
Canberra	1.3	0.824	1.261	0.747	1.366	0.717
Chebyshev	1.335	0.797	1.118	0.756	1.168	0.740
Minkowski	1.204	0.796	1.016	0.733	1.016	0.725
Cosine	1.236	0.758	0.944	0.658	0.971	0.673
Correlation	1.018	0.292	0.903	0.259	0.882	0.252
Hamming	1.406	0.489	1.216	0.425	1.303	0.445
Manhattan	1.191	0.495	0.994	0.743	1.133	0.717
All-combine	9.672	6.587	8.228	5.492	8.572	5.707

Table 5. Training and predicting time in seconds for the Wisconsin diagnostic breast cancer (WDBC) .

KNN model distances	Training time WOFS(s)	Predicting time WOFS(s)	Training time WLSVC(s)	Predicting time WLSVC(s)	Training time WCHI2(s)	Predicting time WCHI2(s)
Euclidean	1.122	0.845	1.019	0.729	1.055	0.816
Canberra	1.378	0.879	1.098	0.715	1.172	0.808
Chebyshev	1.148	0.834	1.106	0.721	1.138	0.818
Minkowski	1.212	0.843	1.022	0.714	1.191	0.837
Cosine	0.943	0.748	0.886	0.627	0.896	0.745
Correlation	0.938	0.719	0.813	0.604	0.905	0.675
Hamming	1.219	0.439	0.948	0.677	1.004	0.781
Manhattan	1.110	0.841	0.98	0.701	1.01	0.759
All-combine	8.665	6.011	7.550	5.115	7.912	5.645

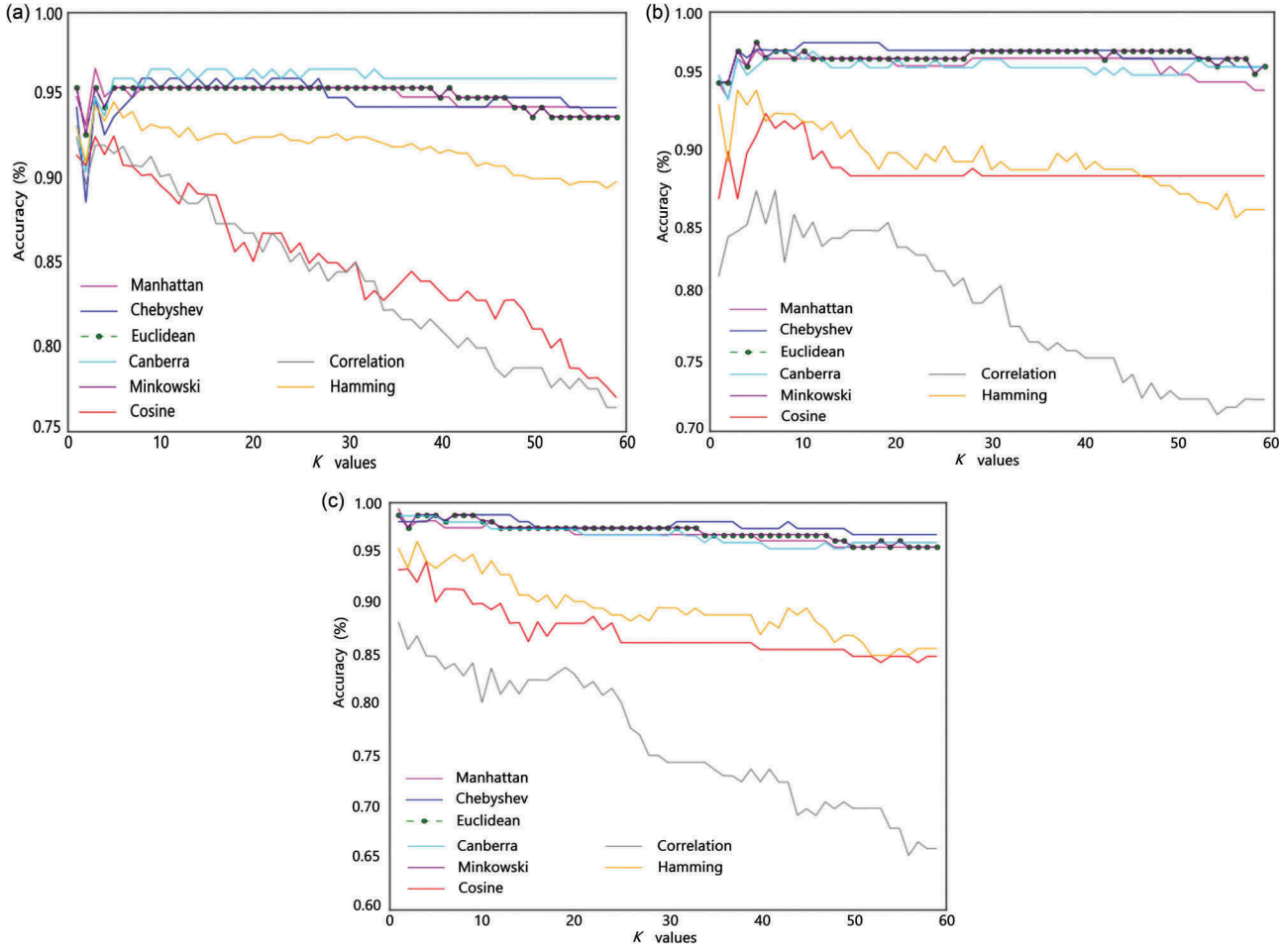

Figure 4. (a) Accuracy of distance functions for WBC by using WOFS. (b) Accuracy of distance functions for WBC by using WLSVC. (c) Accuracy of distance functions for WBC by using WCHI2.

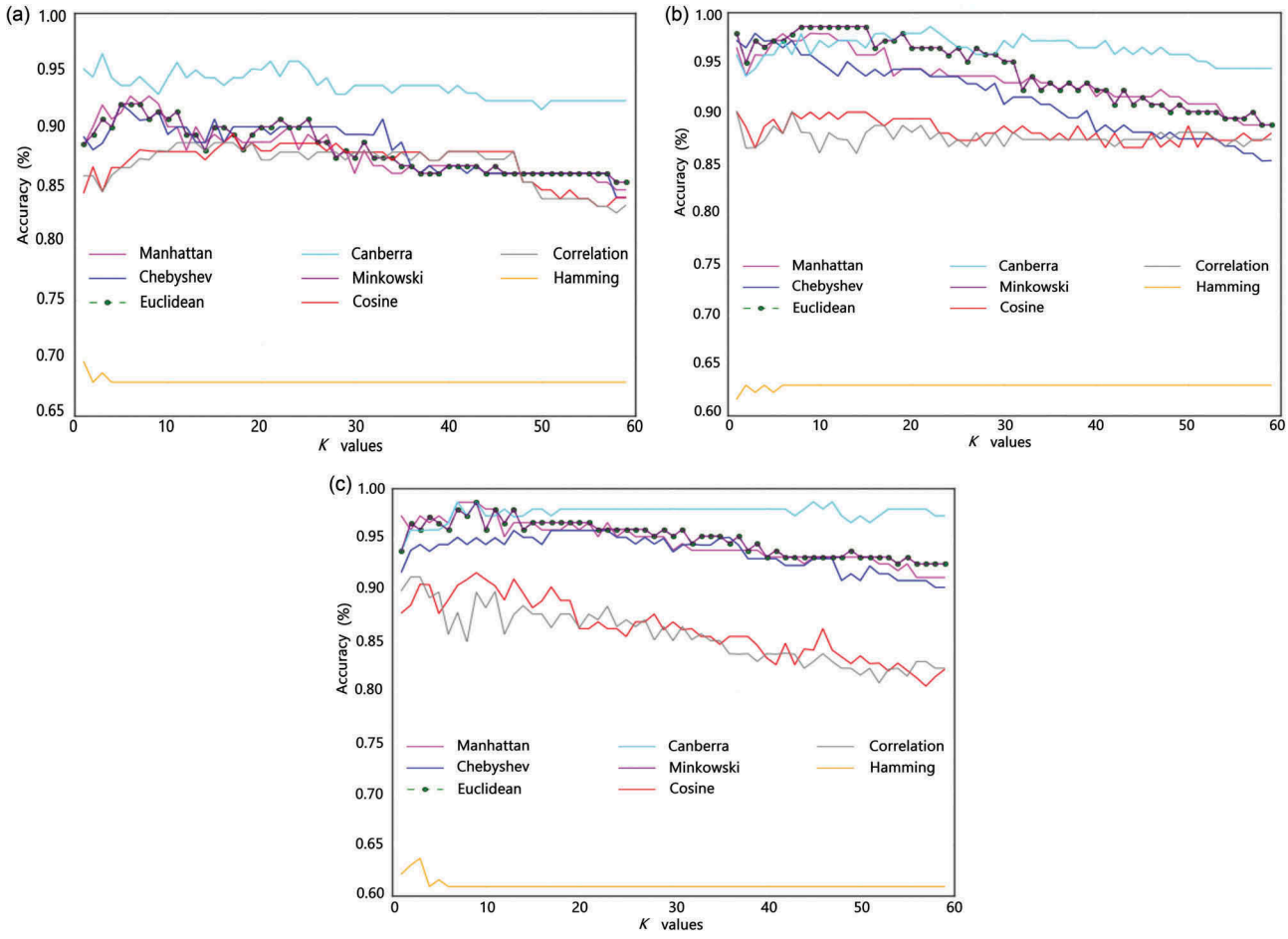

Figure 5. (a) Accuracy of distance functions for WDBC by using WOFS. (b) Accuracy of distance functions for WDBC by using WLSVC. (c) Accuracy of distance functions for WDBC by using WCHI2.

$$Chabyshev_{d(a,b)} = \lim_{p\to\infty}\left(\sum_{i=1}^{n}|a_i - b_i|^p\right)^{\frac{1}{p}} = max_{i=1}^{n}|a_i - b_i|. \tag{11}$$

3.4. Performance evaluation metrics

In order to perform the evaluation of our classifiers, different performance metrics have been considered. These metrics are obtained with the help of the so-called confusion matrix used for the classification tasks. Such evaluation metrics are also widely used in breast cancer classification as previously discussed in Sahu, Mohanty, and Rout (2019); Islam et al. (2018). The terminologies related to the confusion matrix are as follows. TP is true positive, which means the outcome is predicted as positive, and the results are also positive or true. FP is false positive, which shows that the result is actually negative, but the system predicts it as positive. TN is true negative, in which the prediction is negative, and the result is also negative. Lastly, FN is false negative, where the result is positive, and the prediction is negative. The evaluation indices considered in this study are:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN}\text{x}100\%, \tag{12}$$

$$Sensitivity\ or\ Recall(TPR) = \frac{TP}{TP + FN}\ \text{x}\ 100\% = (1 - FNR)\ \text{x}\ 100\%, \tag{13}$$

$$Specificity, Selectivity(TNR) = \frac{TN}{TN + FP}\ \text{x}\ 100\% = (1 - FPR)\ \text{x}\ 100\%, \tag{14}$$

$$Fall - out(FPR) = \frac{FP}{FP + TN}\ \text{x}\ 100\% = (1 - TNR)\ \text{x}\ 100\%, \tag{15}$$

$$Miss - rate(FNR) = \frac{FN}{FN + TP}\ \text{x}\ 100\% = (1 - TPR)\ \text{x}\ 100\%, \tag{16}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}\ \text{x}\ 100\% = \frac{2TP}{2TP + FP + FN}\text{x}100\%, \tag{17}$$

$$ErrorRate = (1 - Acc)\text{x}100\% = \frac{FP + FN}{TP + TN + FP + FN}\ \text{x}\ 100\%, \tag{18}$$

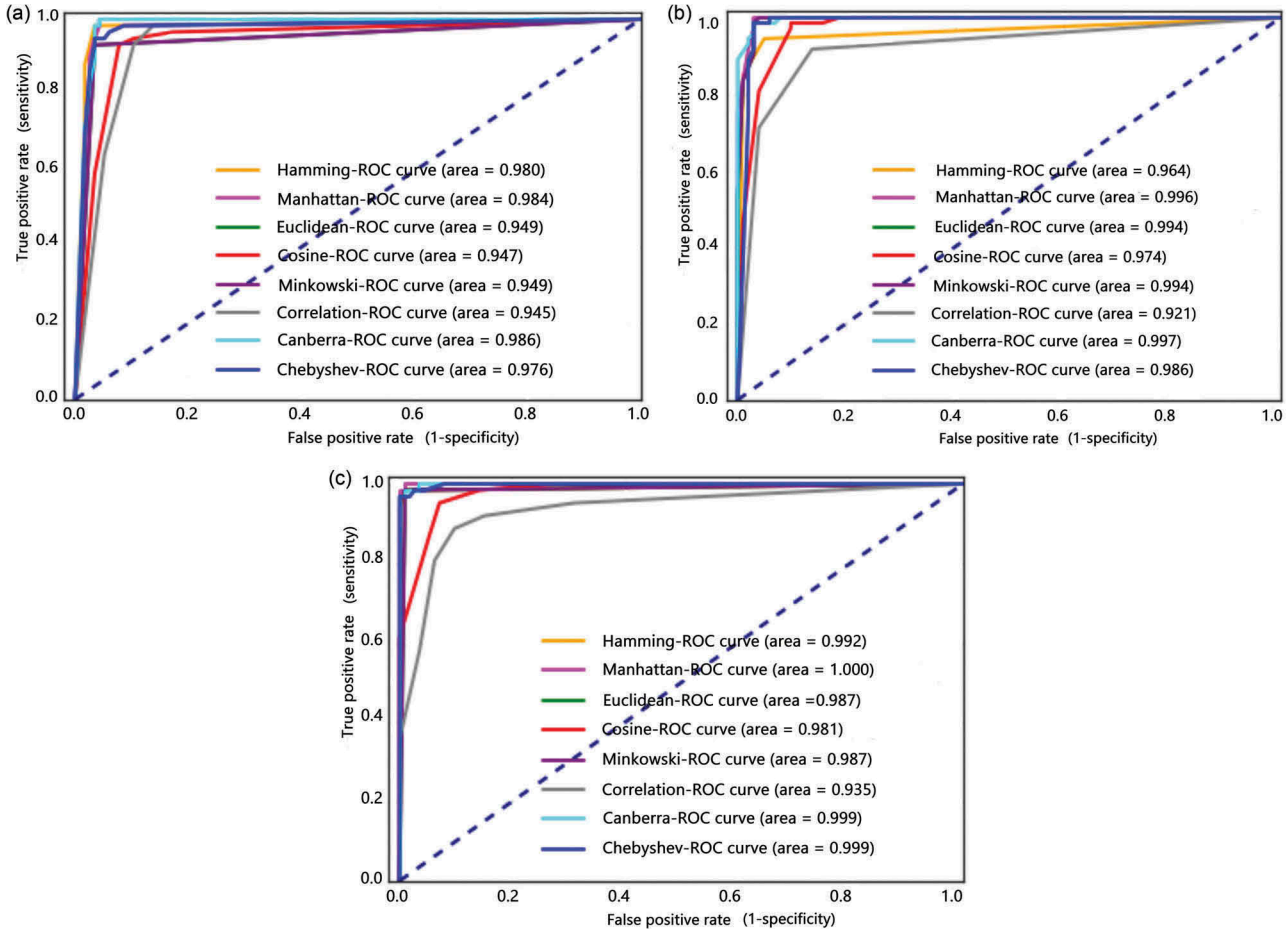

Figure 6. (a) ROC curves with AUC of distance functions for WBC by using WOFS. (b) ROC curves with AUC of distance functions for WBC by using WLSVC. (c) ROC curves with AUC of distance functions for WBC by using WCHI2.

$$\textit{Matthews Correlation Coefficient}(MCC) = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \text{ x } 100\%. \quad (19)$$

Another measurement is the ROC curve with the AUC. It is considered to evaluate the performance of the classifiers. ROC is a graph, which depends upon two parameters: the true positive rate (TPR) as the y-axis and the false positive rate (FPR) as the x-axis. The ROC curve belongs to a category of probability curve, and AUC shows the degree of separability. Figure 3 describes the ROC curves with AUC values. There are three ROC curves with their AUC values denoted as R3, B2, and G1. Model3 shows a larger AUC value as compared to model2 and model1. The larger value of area under the curve, the better the performance of the model. In Figure 3, the value of R3 is much closer to 1.00. Normally, if the classifier scores under the threshold line score, then that particular model performance is the worst. R3> B2> G1 clearly indicates model3 is much better than model2, and, in the same case, the model2 performance is better when compared to model1.

4. Experimental results and discussion

This study was carried out on a core i7, the seventh-generation system with 16GB ram, 256GB SSD, 2TB HDD, and 4GB Nvidia 1050 GPU. Table 4 shows the training and the prediction time on the WBC dataset with the usage of the KNN algorithm. Eight distinct distance functions are considered with a suitable nearest neighbor K value ranging from 1 to 59. The same experiment was repeated three times on the WBC dataset. The first one was completed without any feature selection denoted as WOFS. After ignoring the sample ID, the number column, and the class column, there were a total of nine features available. The second attempt was the $L1$ based feature selection, and the linear support vector classifier was implemented. This approach is denoted as WLSVC. Six features have been selected in this case. The last one is employed with the Chi-square feature selection denoted as WCHI2, and seven features have been selected from the WBC dataset. It can be expected that a smaller number of features reduces the training and testing time. The lower the training and prediction time, the more efficient the classifier. In the case of time management, the correlation distance function in KNN in all attempts is more efficient compared to other distance functions. The performance of distance functions is also shown in Figures 3, 5 and 7. These figures show accuracy, misclassification error, and ROC curves with AUC values for all distance functions for WBC. Table 5 describes the testing and training time of all distance functions on the WDBC dataset. The number of total features after excluding the ID number and the diagnosis column in WOFS is

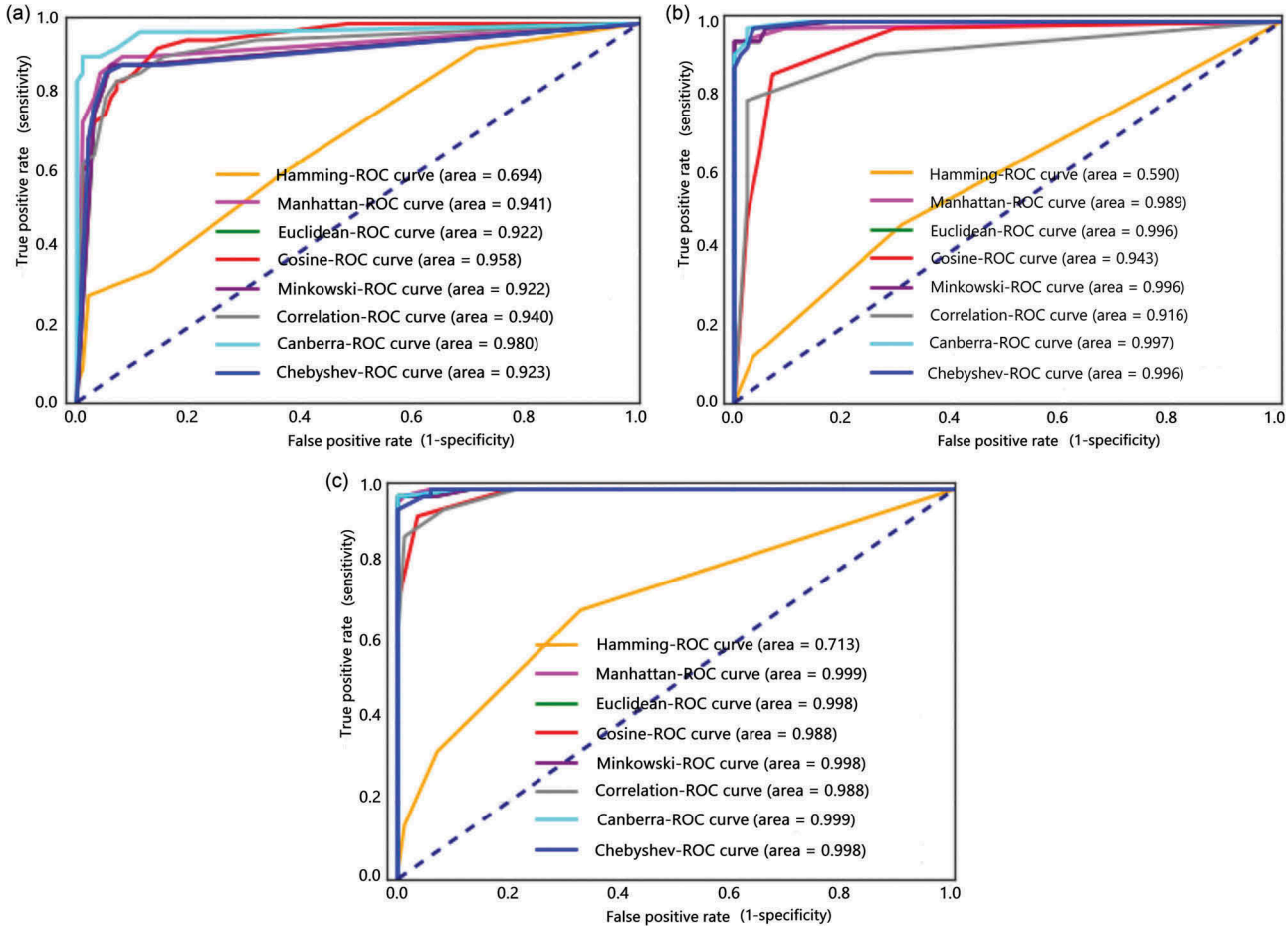

Figure 7. (a) ROC curve with AUC of distance functions for WDBC by using WOFS. (b) ROC curve with AUC of distance functions for WDBC by using WLSVC. (c) ROC curve with AUC of distance functions for WDBC by using WCHI2.

(30). Using the WLSVC, there are (14) features. In the case of WCHI2, (20) features are selected. All distance functions prediction time in WCHI2 is 5.11 seconds, which was clearly a winner in this aspect.

The experimental process involves the WBC and WDBC datasets. Eight different distance functions have been studied by imposing feature selection techniques on each dataset. The results shown in Figures 4 and 6 belong to the WBC dataset. Each figure consists of three subparts. Part (a) involves manipulation of the experimental procedure without the feature selection denoted by WOFS. Part (b) includes the repetition of the same scheme with the $L1$ based selection from the model by using the WLSVC feature selection technique. Part (c) focuses on the Chi-square feature selection technique WCHI2. Table 3 reflects the selection of the features in each mode with each dataset. Figure 4 shows the achieved accuracy rates by the applied algorithm KNN on different distance functions with K values ranging from 1 to 59. Figures 6 and 7 reflect the ROC curve with the AUC values for distance functions, uniformly in the same manner. Figures 5 and 7 belong to the WDBC dataset, which shows the accuracy and ROC curve with AUC values, respectively. The results in Figure 4 clearly indicate the achieved accuracy rates of each distance function with various K values. Figure 4(a) describes the result without the feature selection. Canberra distance performs well with the K value being nine. The highest achieved accuracy is 96.57%, and the lowest accuracy is attained by the correlation distance at the K value as 58, and the accuracy is 73.8%. Figure 4(b) demonstrates the $L1$ based feature selection. In this experiment, the best accuracy gained by the Euclidean and Minkowski distance functions is 97.71% at $K = 5$. Chebyshev also has a 97.71% accuracy with a K value equal to 10. Figure 4(c) shows the Chi-square feature selection, and the overall best accuracy achieved by the Manhattan distance function at $K = 1$ is 99.42%, which is the best among all the techniques implemented on the WBC dataset. Figure 5 narrates the results on the WDBC dataset. Overall, the best accuracy is achieved by Canberra at $K = 8$ and Manhattan with $K = 7$. The obtained result is 98.62% by using the Chi-square-based feature selection methodology. Results with the $L1$ based linear support vector classifier feature selection are also very close to the highest achieved accuracy results. The Euclidean and Minkowski distance function at $K = 8$ achieves 98.59% accuracy. Canberra is also virtually equal to the above result with $K = 22$.

Another important parameter to check the performance of any model related to a classification problem is the ROC curve with the AUC values. The AUC scores range from '0'

Table 6. Performance evaluation metrics for WOFS, WLSVC, WCHI2 on WBC data.

Without feature selection (WOFS)									
Performance Evaluation Metrics		Accuracy %	Recall%	Fall-out %	Specificity %	F1-score%	Error-rate %	MCC %	Miss-rate %
KNN- distance functions with best *K* values range from 1 to 59.	Euclidean $K = 1$	95.42	96.55	6.77	93.22	96.55	4.58	89.77	3.44
	Manhattan $K = 3$	96.00	97.39	6.66	93.22	96.96	4.00	91.09	2.60
	Chebyshev $K = 8$	96.00	97.39	6.66	93.33	96.96	4.00	91.09	2.60
	Minkowski $K = 1$	95.42	96.55	6.77	93.22	96.55	4.58	89.77	3.44
	Cosine $K = 5$	92.57	96.39	14.06	93.22	94.27	7.43	83.88	3.60
	Correlation $K = 1$	90.85	96.29	17.91	82.08	92.85	9.15	80.59	3.70
	Hamming $K = 7$	94.85	94.21	3.70	96.29	96.20	5.15	88.43	5.78
	Canberra $K = 9$	96.57	98.24	6.55	93.44	97.39	3.43	92.42	1.75
Feature selection with *L*1 based linear support vector classifier (WLSVC)									
KNN- distance functions with best *K* values range from 1 to 59.	Euclidean $K = 5$	97.71	99.00	4.00	96.00	98.01	2.29	95.34	1.00
	Manhattan $K = 5$	97.14	98.01	4.05	95.94	97.53	2.86	94.14	1.98
	Chebyshev $K = 10$	97.71	99.00	4.00	96.00	98.01	2.29	95.34	1.00
	Minkowski $K = 5$	97.71	99.00	4.00	96.00	98.01	2.29	95.34	1.00
	Cosine $K = 6$	92.57	96.84	12.5	87.5	93.40	7.43	85.20	3.15
	Correlation $K = 13$	85.71	82.35	7.14	92.85	88.68	14.29	71.15	17.64
	Hamming $K = 1$	93.14	90.17	1.58	98.41	94.39	6.86	86.24	9.82
	Canberra $K = 7$	97.14	98.98	5.26	94.73	97.51	2.86	94.21	1.01
Feature selection with Chi2 (WCHI2)									
KNN- distance functions with best *K* Values range from 1 to 59.	Euclidean $K = 3$	98.85	99.11	1.59	98.38	99.12	1.15	97.50	0.88
	Manhattan $K = 1$	99.42	99.12	0.00	100	99.56	0.58	98.75	0.87
	Chebyshev $K = 5$	98.85	98.26	1.60	100	99.12	1.15	97.51	1.73
	Minkowski $K = 3$	98.85	99.11	1.59	98.38	99.12	1.15	97.50	0.88
	Cosine $K = 4$	94.86	97.27	9.23	90.76	95.96	5.14	88.94	2.72
	Correlation $K = 5$	89.14	89.83	12.28	87.71	91.77	10.86	64.66	10.16
	Hamming $K = 3$	96.00	94.16	0.00	100	96.99	4.00	91.39	5.83
	Canberra $K = 1$	98.85	99.11	1.60	98.38	99.12	1.15	97.50	0.88

Table 7. Performance evaluation metrics for WOFS, WLSVC, WCHI2 on WDBC data.

Without feature selection (WOFS)									
Performance evaluation metrics		Accuracy %	Recall%	Fall-out%	Specificity%	F1score%	Error-rate%	MCC%	Miss-rate%
KNN- distance functions with best *K* values range from 1 to 59.	Euclidean $K = 5$	92.30	94.79	12.76	87.23	94.30	7.70	82.48	5.20
	Manhattan $K = 6$	93.00	93.94	9.09	90.90	94.89	7.00	83.83	6.06
	Chebyshev $K = 5$	92.30	93.87	11.11	88.88	94.35	7.70	82.28	6.12
	Minkowski $K = 5$	92.30	94.79	12.76	87.23	94.30	7.70	82.48	5.20
	Cosine $K = 17$	89.50	91.83	15.55	84.44	92.30	10.5	75.83	8.16
	Correlation $K = 8$	88.90	90.09	14.28	85.71	91.91	11.1	73.92	9.90
	Hamming $K = 1$	69.90	69.56	20.00	80.00	81.70	30.1	19.49	30.43
	Canberra $K = 3$	96.50	96.00	2.30	97.67	97.46	3.50	91.95	4.00
Feature selection with *L*1 based linear support vector classifier (WLSVC)									
KNN- distance functions with best *K* values range from 1 to 59.	Euclidean $K = 8$	98.59	100.0	4.65	95.34	99.00	1.41	96.68	0.00
	Manhattan $K = 6$	97.90	99.00	4.76	95.23	98.52	2.10	95.39	0.99
	Chebyshev $K = 9$	96.50	98.01	7.14	92.85	97.53	3.50	91.52	1.98
	Minkowski $K = 8$	98.59	100.0	4.65	95.34	99.00	1.41	96.68	0.00
	Cosine $K = 1$	90.20	90.80	10.71	89.28	91.86	9.80	79.61	9.19
	Correlation $K = 1$	90.20	87.36	4.16	95.83	92.22	9.80	80.01	12.63
	Hamming $K = 2$	63.03	63.12	50.00	50.00	77.05	36.97	3.190	36.87
	Canberra $K = 22$	98.59	100.0	4.65	95.34	99.00	1.41	96.68	0.00
Feature selection with Chi2 (WCHI2)									
KNN- distance functions with best *K* values range from 1 to 59.	Euclidean $K = 9$	98.60	97.84	0.00	100.0	98.91	1.40	96.99	2.15
	Manhattan $K = 7$	98.62	97.89	0.00	100.0	98.93	1.38	96.94	2.10
	Chebyshev $K = 13$	95.80	97.89	8.33	91.66	96.87	4.20	90.53	2.10
	Minkowski $K = 9$	98.60	97.84	0.00	100.0	98.91	1.40	96.99	2.15
	Cosine $K = 9$	91.60	94.18	12.28	87.71	93.10	8.40	82.42	5.81
	Correlation $K = 2$	91.00	89.47	6.25	93.75	92.89	9.00	80.78	10.52
	Hamming $K = 3$	63.66	62.85	0.00	100.0	77.19	36.34	18.51	37.14
	Canberra $K = 8$	98.62	97.89	0.00	100.0	98.93	1.38	96.94	2.10

to '1,' and the closer the value is to '1,' the better the performance of the classifier. Figures 6 and 7 describe this metric for WBC and WDBC, respectively. The highest AUC value is achieved by the Manhattan distance function with a value of '1.00' by using the Chi-square feature selection on the WBC dataset, as per the results shown in Figure 6 (c). Figure 7 narrates the AUC values on the WDBC dataset. In spite of the Euclidean and Minkowski distance functions' popularity for KNN, the Canberra and Manhattan distance functions are clearly the winners for the WBC datasets. In Figure 7(c), both the Manhattan and Canberra distance functions have AUC values of 0.999 after using the Chi-

Table 8. Comparison of this study with other existing models.

[References]	Methodology	Accuracy %	Data set
(Alkhasawneh and Tay 2018)	CFNN	97.70	Wisconsin breast cancer (WBC)
	HECFNN	97.12	
	ENN	88.20	
(Marcano-Cedeño, Quintanilla-Domínguez, and Andina 2011)	AMMLP	99.26	Wisconsin breast cancer (WBC)
(Karabatak 2015)	Weighted-NB	98.54	Wisconsin breast cancer (WBC)
(Kong et al. 2016)	JSDA	93.85	Wisconsin diagnostic breast cancer (WDBC)
(Mert et al. 2015)	40% test, *k*-NN	92.56	Wisconsin diagnostic breast cancer (WDBC)
	1 feature reduced by ICA		
(Aalaei et al. 2016)	PSO-FS	96.9	
	GA-FS	96.6	
	ANN-FS	96.7	
		97.2	Wisconsin diagnostic breast cancer (WDBC)
		96.6	
		97.3	
(Sheikhpour, Sarram, and Sheikhpour 2016)	PSO-KDE	98.53	Wisconsin breast cancer (WBC)
	GA-KDE	98.53	
		98.45	Wisconsin diagnostic breast cancer (WDBC)
		98.45	
(Nilashi et al. 2017)	EM-PCA-CART-Fuzzy Rule-Based	94.1	Wisconsin diagnostic breast cancer (WDBC)
(Chidambaranathan 2016)	K-SVM	97.38	Wisconsin diagnostic breast cancer (WDBC)
(Osman 2017)	Two step SVM	99.10	Wisconsin breast cancer (WBC)
This Study	KNN-Manhattan ($K = 1$) Chi2 FS	99.42	Wisconsin breast cancer (WBC)
	KNN-Canberra ($K = 8$)	98.62	Wisconsin diagnostic breast cancer (WDBC)
	KNN- Manhattan ($K = 7$)		
	Chi2 FS		

Table 9. Comparison of this research with another relevant approach.

Wisconsin breast cancer (WBC)		
	Highest achieved accuracy with *K* value	
KNN distance functions	Reference (Medjahed, Saadi, and Benyettou 2013)	This study
Euclidean	$K = 1$, 98.70%	$K = 3$, 98.85%
Manhattan	$K = 1$, 98.48%	$K = 1$, 99.42%
Cosine	$K = 1$, 95.67%	$K = 4$, 94.86%
Correlation	$K = 10$, 95.35%	$K = 5$, 89.16%

square feature selection. The Hamming distance function, using the *L*1 based linear support vector classifier feature selection technique, shows a poor performance with the AUC value of 0.590. The results show the Manhattan and Canberra distance functions with the Chi-square-based feature selection performs the best for the WBC datasets.

Tables 6 and 7 show the detailed description of the performance assessment metrics for the WBC and the WDBC KNN-based distance functions with their highest accuracy *K* values also shown. Such metrics are based on the confusion matrix. The bold digits show the high achiever KNN-based distance function in each dataset. In the case of the WBC, Manhattan, and the WDBC, both the Canberra and Manhattan distance obtain the best results. For the WBC datasets, Canberra and Manhattan show much better results compared with other distance functions.

5. Comparison with other existing studies

The approach proposed in this paper has been compared with other existing models in Table 8. The results show that our models have achieved a distinctive development by only using the simple KNN with the most suitable distance function and a suitable *K* value by using the Chi-square-based feature selection technique. Table 9 compares this research with another identical study. Our approach achieves a higher accuracy rate, lower training and prediction time, as shown in Tables 8 and 9. Other existing models used complex models with a fusion of various algorithms, and a few focused-on hybrid approaches in combination with feature engineering to achieve the best results. While such phenomena can lead to good accuracy, it increases the time for training and validation. The methodology in this manuscript is the simplest and most cost-effective way of acquiring the highest accuracy in a small interval of time for training and testing. The study in Table 9 involves an equivalent approach to the WBC by using only four distance functions in comparison with our study, which includes eight distance functions on the WBC and WDBC. The best accuracy on the WBC was achieved by the Manhattan distance function with a *K* value equal to one, as clearly shown in Table 8.

6. Conclusions

This paper is based on one of the most famous classifiers, KNN. Eight different distance functions with their optimal *K* value range from 1 to 59 were implemented on the WBC and the WDBC datasets. The approaches proposed in this study involved feature engineering with means assigned to missing values. Various performance evaluation metrics were also used on both datasets. Experiments included

WOFS, with $L1$ based LSVC feature selection and the Chi-square-based feature selection. Our study shows a simple and robust idea to achieve a higher accuracy result by using KNN. A good K value and an approach of selecting the most suitable distance function with the Chi-square feature selection can obtain remarkable results. In the WBC dataset, the Manhattan distance function with $K = 1$ achieved a 99.42% accuracy and 1.00 AUC value by using the Chi-square feature selection. For the WDBC dataset, both the Canberra and the Manhattan distance functions gained the best result of 98.62% with a K value of 7 or 8 after adopting the Chi-square feature selection technique. The ROC based AUC values for both distance functions were 0.999. Comparison of this study with other more complex existing models clearly revealed the robustness and effectiveness of this approach. Our results show that the Chi-square feature selection with the Canberra or Manhattan distance function having a K value range from 1 to 9 was a good classification technique for the WBC and WDBC datasets. By using the same methodology, the KNN algorithm can do better than any other time-consuming and complex algorithm in breast cancer classification problems.

Nomenclature

K	number of neighboring elements
$L1$	$L1$-norm
X	An attribute
X_n	total number of instances
X_p	number of positive instances, with feature X
X_{neg}	number of negative instances, with feature X
$X_{p\_}$	number of positive instances, without feature X
$X_{neg\_}$	number of negative instances, without feature X
X_{pn}	number of instances, with feature X
X_{nn}	number of instances, without feature X
σ	standard deviation
a, b	feature vectors
cov	covariance
c	Pearson correlation coefficient
p	Cartesian coordinate point
d	distance

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Aalaei, S., H. Shahraki, A. Rowhanimanesh, and S. Eslami. 2016. "Feature Selection Using Genetic Algorithm for Breast Cancer Diagnosis: Experiment on Three Different Datasets." *Iranian Journal of Basic Medical Sciences* 19 (5): 476–482. doi:10.1063/1.108412.

Abdel-Ilah, L., and H. Šahinbegovi. 2017. "Using Machine Learning Tool in Classification of Breast Cancer." *CMBEBIH 2017, IFMBE Proceedings* 62 (1): 3–8. doi:10.1007/978-981-10-4166-2.

Abdel-Zaher, A. M., and A. M. Eldeib. 2016. "Breast Cancer Classification Using Deep Belief Networks." *Expert Systems with Applications* 46 (2): 139–144. doi:10.1016/j.eswa.2015.10.015.

Alkhasawneh, M. S., and L. T. Tay. 2018. "A Hybrid Intelligent System Integrating the Cascade Forward Neural Network with Elman Neural Network." *Arabian Journal for Science & Engineering* 43 (12): 6737–6749. doi:10.1007/s13369-017-2833-3.

Asri, H., H. Mousannif, H. Al Moatassime, and T. Noel. 2016. "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis." *Procedia Computer Science* 83 (Fams): 1064–1069. doi:10.1016/j.procs.2016.04.224.

Bihis, M., and S. Roychowdhury. 2015. "A Generalized Flow for Multi-Class and Binary Classification Tasks: An Azure ML Approach." *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data*, Santa Clara, CA. USA, 29 Oct-1 Nov.2015: 1728–1737 New York: IEEE.

Chen, H. L., B. Yang, J. Liu, and D. Y. Liu. 2011. "A Support Vector Machine Classifier with Rough Set-Based Feature Selection for Breast Cancer Diagnosis." *Expert Systems with Applications* 38 (7): 9014–9022. doi:10.1016/j.eswa.2011.01.120.

Chidambaranathan, S. 2016. "Breast Cancer Diagnosis Based on Feature Extraction by Hybrid of K-Means and Extreme Learning Machine Algorithms." *ARPN Journal of Engineering and Applied Sciences* 11 (7): 4581–4586. doi:10.1016/j.eswa.2013.08.044.

Hall, M. A., and G. Holmes. 2003. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining." *IEEE Transactions on Knowledge and Data Engineering* 15 (6): 1437–1447. doi:10.1109/TKDE.2003.1245283.

Hu, L. Y., M. W. Huang, S. W. Ke, and C. F. Tsai. 2016. "The Distance Function Effect on K- Nearest Neighbor Classification for Medical Datasets." *Springer Plus* 5 (1): 1304. doi:10.1186/s40064-016-2941-7.

Institute, National Cancer for Surveillence, Epidemology and End Result Program. 2018. "Female Breast Cancer - Cancer Stat Facts." National Cancer Institute. Accessed Feburary 3 2019. https://seer.cancer.gov/statfacts/html/breast.html

Islam, M. M., H. Iqbal, M. R. Haque, and M. K. Hasan. 2018. "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors." *5th IEEE Region 10 Humanitarian Technology Conference* 2017, R10-HTC, Dhaka, Bangladesh, 21–23 Dec 2017: 226–229. New York: IEEE.

Juneja, K., and C. Rana. 2018. "An Improved Weighted Decision Tree Approach for Breast Cancer Prediction." *International Journal of Information Technology* 1–8. doi:10.1007/s41870-018-0184-2.

Karabatak, M. 2015. "A New Classifier for Breast Cancer Detection Based on Naïve Bayesian." *Measurement: Journal of the International Measurement Confederation* 72: 32–36. doi:10.1016/j.measurement.2015.04.028.

Kong, H., Z. Lai, X. Wang, and F. Liu. 2016. "Breast Cancer Discriminant Feature Analysis for Diagnosis via Jointly Sparse Learning." *Neurocomputing* 177: 198–205. doi:10.1016/j.neucom.2015.11.033.

Li, C., S. Zhang, H. Zhang, L. Pang, K. Lam, C. Hui, and S. Zhang. 2012. "Using the K-Nearest Neighbor Algorithm for the Classification of Lymph Node Metastasis in Gastric Cancer." *Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine* 2012 (4): 876545. doi:10.1155/2012/876545.

Marcano-Cedeño, A., J. Quintanilla-Domínguez, and D. Andina. 2011. "WBCD Breast Cancer Database Classification Applying Artificial Metaplasticity Neural Network." *Expert Systems with Applications* 38 (8): 9573–9579. doi:10.1016/j.eswa.2011.01.167.

Medjahed, S. A., T. A. Saadi, and A. Benyettou. 2013. "Breast Cancer Diagnosis by Using K-Nearest Neighbor with Different Distances and Classification Rules." *International Journal of Computer Applications* 62 (1): 1–5. doi:10.5120/10041-4635.

Mert, A., N. Kiliç, E. Bilgili, and A. Akan. 2015. "Breast Cancer Detection with Reduced Feature Set." *Computational and Mathematical Methods in Medicine* 2015: 1–11. doi:10.1155/2015/265138.

Nilashi, M., O. Ibrahim, H. Ahmadi, and L. Shahmoradi. 2017. "A Knowledge-Based System for Breast Cancer Classification Using Fuzzy Logic Method." *Telematics and Informatics* 34 (4): 133–144. doi:10.1016/j.tele.2017.01.007.

Osman, A. H. 2017. "An Enhanced Breast Cancer Diagnosis Scheme Based on Two-Step-SVM Technique." *International Journal of Advanced Computer Science & Applications* 8 (4): 158–165. doi:10.14569/ijacsa.2017.080423.

Osmanović, A., S. Halilović, L. A. Ilah, A. Fojnica, and Z. Gromilić. 2019. "Machine Learning Techniques for Classification of Breast Cancer." In *World Congress on Medical Physics and Biomedical Engineering*

*2018,1*IFMBE proceedings:edited by L. Lhotska, L. Sukupova, I. Lacković, and G. Ibbott. Vol. 68/1,,197–200. Berlin: Singapore: Springer.

Park, K., A. Ali, D. Kim, Y. An, M. Kim, and H. Shin. 2013. "Robust Predictive Model for Evaluating Breast Cancer Survivability." *Engineering Applications of Artificial Intelligence* 26 (9): 2194–2205. doi:10.1016/j.engappai.2013.06.013.

Pedregosa, F., R. Weiss, and M. Brucher. 2011. "Scikit-Learn : Machine Learning in Python." *Journal Of Machine Learning Research* 12 (2011): 2825–2830. doi:10.1016/j.patcog.2011.04.006.

Pérez, N., R. R. Frias, and A. Silva. 2015. "Ensemble Features Selection Method as Tool for Breast Cancer Classification Isabel Ramos." *International Journal of Image Mining* 1 (2–3): 224–244. doi:10.1504/ijim.2015.073019.

Rathi, M. 2016. "Hybrid Approach to Predict Breast Cancer Using Machine Learning Techniques." *International Journal of Computer Science Engineering (IJCSE)* 5 (03): 125–136.

Sahu, B., S. N. Mohanty, and S. K. Rout. 2019. "A Hybrid Approach for Breast Cancer Classification and Diagnosis." *EAI Endorsed Transactions on Scalable Information Systems* 6 (20): 1–8. doi:10.4108/eai.19-12-2018.156086.

Salama, G. I., M. B. Abdelhalim, and M. A. Zeid. 2012. "Breast Cancer Diagnosis on Three Different Datasets Using Multi-classifiers." *International Journal of Computer and Information Technology* 01 (01): 36–43. www.ijcit.com

Sheikhpour, R., M. A. Sarram, and R. Sheikhpour. 2016. "Particle Swarm Optimization for Bandwidth Determination and Feature Selection of Kernel Density Estimation Based Classifiers in Diagnosis of Breast Cancer." *Applied Soft Computing Journal* 40: 113–131. doi:10.1016/j.asoc.2015.10.005.

Tripathy, R. K., S. Mahanta, and S. Paul. 2014. "Artificial Intelligence-Based Classification of Breast Cancer Using Cellular Images." *Royal Society of Chemistry* 2014 (18): 9349–9355. doi:10.1039/c3ra47489e.

Wolberg, W. H., W. N. Street, and O. L. Mangasarian. 2019. *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) & Breast Cancer Wisconsin (Original) Data Sets*. University of California, Irvine School of Information & Computer Sciences: Accessed February 7 2019. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

You, H., and G. Rumbe. 2010. "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data." *International Journal of Interactive Multimedia and Artificial Intelligence* 1 (3): 5–12. doi:10.9781/ijimai.2010.131.