DE GRUYTER
OPEN

Research Paper

Detection of Malignant and Benign Breast Cancer Using the ANOVA-BOOTSTRAP-SVM

Borislava Petrova Vrigazova

Department of Statistics and Econometric, Faculty of Economics and Business Administration, Sofia University, Bulgaria

Citation: Vrigazova, Borislava Petrova. "Detection of malignant and benign breast cancer using the ANOVA-BOOTSTRAP-SVM." *Journal of Data and Information Science*, vol. 5, no. 2, 2020, pp. 62–75. https://doi.org/10.2478/jdis-2020-0012

Received: Dec. 22, 2019
Revised: Mar. 18, 2020
Accepted: Apr. 7, 2020

Abstract

Purpose: The aim of this research is to propose a modification of the ANOVA-SVM method that can increase accuracy when detecting benign and malignant breast cancer.

Methodology: We proposed a new method ANOVA-BOOTSTRAP-SVM. It involves applying the analysis of variance (ANOVA) to support vector machines (SVM) but we use the bootstrap instead of cross validation as a train/test splitting procedure. We have tuned the kernel and the C parameter and tested our algorithm on a set of breast cancer datasets.

Findings: By using the new method proposed, we succeeded in improving accuracy ranging from 4.5 percentage points to 8 percentage points depending on the dataset.

Research limitations: The algorithm is sensitive to the type of kernel and value of the optimization parameter C.

Practical implications: We believe that the ANOVA-BOOTSTRAP-SVM can be used not only to recognize the type of breast cancer but also for broader research in all types of cancer.

Originality/value: Our findings are important as the algorithm can detect various types of cancer with higher accuracy compared to standard versions of the Support Vector Machines.

Keywords Breast cancer detection; ANOVA; Bootstrap; Support vector machines

1 Introduction

Breast cancer has been one of the most common types of cancer among women. It can be benign and malignant. Medical researchers have been focused on timely detection of malignant cancer as it spreads in surrounding tissues in the body (Siegel, Miller, & Jemal, 2015). Timely detection of malignant cancer is important for the survival rate after undergoing cancer treatment. Some recent research show that the breast cancer survival rate in the developed countries is between 80% and 90% (Mustafa et al., 2016), while it is much lower in developing countries (Jemal et al.,

† Corresponding author: Borislava Petrova Vrigazova (E-mail: vrigazova@uni-sofia.bg).

2011). To keep the survival rate increasing, interdisciplinary studies have developed both medical tests (Breit et al., 2019; Noske et al., 2020) and machine learning algorithms (Singh, 2019; Wu et al., 2019) to detect malignant cancer cells. Medical tests focus on detecting deviations in breast cells, while machine learning algorithms use image classification (Wang et al., 2020) for screening deviations in breast tissue. Machine learning classification algorithms (Salama, Abdelhalim, & Zeid, 2012; Wang et al., 2019) are also used to detect malignant breast cancer. Most of them (Asri et al., 2016) achieve accuracy lower than 99%.

The aim of this study is to propose a modification of the SVMs, which classifies malignant tumors with 99.6% accuracy. This result is an important contribution to academic literature as this is the highest accuracy achieved on the Wisconsin breast cancer dataset using classification algorithms, particularly the SVMs. Also, our proposition results in the smallest error rate compared to other existing methods (Asri et al., 2016; Maldonado et al., 2014; Vrigazova and Ivanov, 2019). We also tested the algorithm on the mammographic mass breast cancer dataset (Elter, Schulz-Wendtland, & Wittenberg, 2007) and it boosted the classification of benign and malignant cancer significantly compared to the cassic SVM. Therefore, our algorithm can be applied to various datasets to boost chances of discovering malignant breast cancer.

2 Literature review

Deep learning methods for detecting breast cancer include neural networks (Yan et al., 2019; Ting, Tan, & Sim, 2019) for image classification of affected breast tissue. Researchers in this field use images of cancerous breast tissue to build classification algorithms that can predict the stage of development of breast cancer. Neural networks can be combined with feature selection like the ridge and liner discriminant analysis to perform image classification (Toğaçar, Ergen, & Cömert, 2020). These algorithms have wide application in medical screening as they can classify a patient as healthy or sick and determine the type of breast cancer without requiring prior knowledge about the lack or presence of breast cancer (Ting, Tan, & Sim, 2019). These algorithms also achieve high accuracy when predicting the type of cancer (Ting, Tan, & Sim, 2019; Toğaçar, Ergen, & Cömert, 2020) and they can be used as additional diagnostic method along with medical tests.

Machine learning techniques are used to extract text information about typical symptoms of malignant breast cancer by checking medical records of patients. For instance, Forsyth et al. (2018) built a machine learning algorithm to find the most common symptoms of breast cancer that patients reported. They algorithm checked 103,564 sentences to identify pain, fatigue, and nausea as the most common cancer

symptoms. Zhang et al. (2019) also studied characteristics of breast cancer to predict its occurrence. They achieved f1-score of 93.53% using neural networks. Such research extends medical knowledge about causes and symptoms of breast cancer dataset and increases chances of correct and timely diagnosis.

Along with neural networks, classification methods are another popular machine learning technique to explore breast cancer. Unlike previously mentioned work, classification methods use quantitative data to predict the type of breast cancer. The most common dataset used is the Wisconsin breast cancer dataset. It contains quantitative data about breast cancer physical characteristics, while the target variable is a categorical variable corresponding to benign or malignant cancer. The classification task is to predict the malignant cases. Many authors have examined this dataset using various approaches. For instance, Liu et al. (2019) devised a novel approach for feature selection on the dataset called IGSAGAW-CSSVM to improve prediction accuracy. Asri et al. (2016) applied support vector machines, k-nearest neighbors, naive Bayes and decision tree. He achieved accuracy of 97.13% in predicting the malignant cases. The purpose of classification methods is to help diagnose breast cancer based on its physical characteristics.

The decision tree classifier (Quinlan, 1996) achieved 94.7% accuracy on the Wisconsin breast cancer dataset, while ensemble learning algorithms (Bashir, Qamar, & Khan, 2015) achieved 97.4% accuracy. Although not widely used, a nero-rule based approach (Setiono, 2000) managed to achieve 98.2% accuracy on the dataset. Support vector machines (SVMs) is the most commonly used algorithm to achieve high accuracy when predicting malignant breast cancer (Liu et al., 2019; Wang et al., 2018). Previous research (Chaurasia and Pal, 2017) has showed that the SVMs with RBF kernel is the most suitable algorithm to detect malignant breast cancer. They (Chaurasia and Pal, 2017) achieved accuracy of 96.8%. Other papers also showed that the SVMs may be the most appropriate classification method for detecting malignant cancer (Maldonado et al., 2014).

In a previous research we (Vrigazova & Ivanov, 2019) presented a modification of the SVMs called the ANOVA-SVM-BOOTSTRAP and achieved accuracy of 97.5%. In this paper, we tune the ANOVA-BOOTSTRAP-SVM (Vrigazova & Ivanov, 2019) to increase prediction accuracy on the Wisconsin breast cancer dataset. We show that our version outperforms some current research (Maldonado et al., 2014) and (Khairunnahar et al., 2019). We also show that our algorithm significantly boosts the results achieved on other breast cancer datasets compared to the classical SVMs. With this discovery, we propose an optimized version of the Support vector machines that can be applied on various breast cancer datasets to detect malignant cancer. Section 2 presents the algorithm. Sections 3 and 4 present the results and some discussion.

3 Methodology

We call the proposed algorithm for detecting malignant cancer the ANOVA-BOOTSTRAP-RBF-SVM. This procedure detects the number of features that results in the highest accuracy of the Support vector classifier. We use the bootstrap procedure as model validation procedure. The advantages of our algorithm include fast computation, improved classification ability of the support vector classifier and simplicity of application. We perform our research in Python 3.6 by following the steps below.

Step 1: We first standardize the input data by using the StandardScaler() function in Python and applying equation 1:

$$z_{ij} = \frac{x_{ij} - \mu}{\sigma} \tag{1}$$

Step 2: We normalize (eq. 2) the standardized data from step 1 to have values between 0 and 1. We do that by the MinMaxScaler() function in scikitlearn.

$$n_{ij} = \frac{z_{ij} - \min(z_{ij})}{\max\left(z_{ij}\right) - \min(z_{ij})} \tag{2}$$

Step 3: We split the dataset into training and test set by using the 10-fold bootstrap procedure described in (Vrigazova & Ivanov, 2019).

Step 4: We fix C=32 and use the radial basis function (RBF) kernel for the support vector classifier. We chose C=32 as we ran the ANOVA-BOOTSTRAP-RBF-SVM with a grid of integer values for C between 1 and 32. The value of 32 proved to result in the highest accuracy as shown in the next section. The reason we chose this grid of values is the fact that the Wisconsin breast dataset has 32 variables, including the target variable. So, we introduce new constraint on the classical C-optimization problem for SVMs introduced in (Cortes et al., 1995). Eq. 3–4 show our modification:

$$min\frac{1}{2}w^2 w + C\sum_{i=1}^{l}\xi_{i*},\ where\ C > 0\ and\ C = p + 1 \tag{3}$$

subject to

$$y_{i*}\left(w^T\phi(x_i\right) + b1 - \xi_{i*},\ where\ \xi_{i*}0, i = 1\ldots1 \tag{4}$$

The vector w denotes the weight for each variable (p), l denotes the length of the training set, which unlike the standard SVM, we chose using the bootstrap procedure. The term ξ_{i*} marks the error term, while $\phi(x_i)$ denotes the kernel function. In the standard SVM, C is regularization term that shrinks the weight of coefficients. It is a positive constant that is usually determined by preliminary setting a grid of values. The value that provides the highest accuracy is the one used in the model.

In academic literature, there is no existing rule to choose the values in the grid. Therefore, in the standard SVMs the constraint imposed on the regularization term is to be positive. However, we introduce additional constraint, particularly C=1+p, where p is the number of features. The value of 1 corresponds to the fact that the target variable is only one. We derived this rule based on empirical results using the grid [1,1+p].

Step 5: For each percentile of features we run the support vector classifier subject to the constraints introduced in steps 1–4. We use the same Python functions as described in our previpus work (Vrigazova & Ivanov, 2019).

Step 6: We select the combination of variables that produced the highest accuracy.

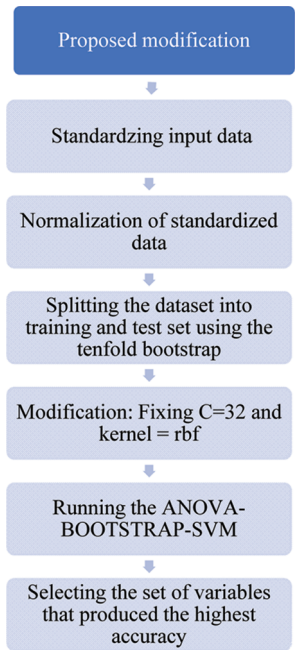

Figure 1. The ANOVA-BOOTSTRAP-RBF-SVM: illustration.

We compare the performance of our modification to other existing methods in academic literature. Next section shows our empirical results and compares them to existing research.

4 Empirical results

Various research has worked on the Wisconsin diagnostic breast dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)). For example, Maldonado et al. (2014) used the mixed linear integer approach to

modify the SVM for the breast cancer dataset. Table 1 shows their results. They achieved the highest accuracy (98.1%) using the MILP1 approach and 26 features. Their approach led to the lowest error rate of 1.9%. The rest of their modifications provided similar results—accuracy of 97.9% and error rate of 2.1%.

Table 1. Maldonado et al.'s results on WBCD (Mixed linear integer approach).

	ACC	AUC	k	Error rate
l2-SVM*	97.900	97.300	31	2.1
LP-SVM*	97.200	96.500	31	2.8
l1-SVM*	97.500	97.200	31	2.5
Fisher+ SVM*	97.900	97.300	31	2.1
RFE-SVM*	97.900	97.300	23	2.1
l0-SVM*	97.900	97.300	16	2.1
MILP1*	98.100	97.700	26	1.9
MILP2*	97.900	97.300	17	2.1

Source: Maldonado et al., 2014; error rate—author's calculations.

In 2019, Khairunnahar et al. (2019) used sigmoid function to improve classification ability on the breast cancer dataset. She achieved highest accuracy of 97.3%, which is comparable to Maldonado's results. We calculated the error rate of her study as Table 2 shows. The lowest error rate that the sigmoid classification produced was 2.6%. However, the AUC score in this case was 90, which was by 7.7 percentage points lower than Maldonado's research. Unlike Maldonado et al., she used fewer features to get this accuracy.

Table 2. Khairunnahar et al.'s results on WBCD (sigmoid classification function).

	ACC	AUC	k	Error rate
Classical system	95.708	82.000	12	4.3
Proposed sigmoid	97.425	90.000	12	2.6
Classical system	95.423	97.540	31	4.6
Proposed sigmoid	96.831	99.000	31	3.2

Source: Khairunnahar et al., 2019, error rate—author's calculations.

Table 3 shows our experimental results on the WBCD that we first obtained in (Vrigazova & Ivanov, 2019).

We manage to achieve accuracy of 97.3% in our previous work (Vrigazova & Ivanov, 2019). We achieved that by using the ANOVA-BOOTSTRAP-L-SVM with C=1 and linear kernel. In the current research we did experiments comparing the performance of the classical SVM to the ANOVA-BOOTSTRAP-L-SVM. We ran experiments with C=1, kernel linear and rbf respectively. In another experiment we changed C=1 to be C=0.1. Table 3 shows that these classical versions of the SVM

Table 3. Vrigazova's results on the WBCD.

	ACC	AUC	k	Error rate
ANOVA-CV-RBF-SVM	97.310	99.750	3	2.7
ANOVA-Bootstrap-L-SVM*	97.270	99.131	24	2.7
ANOVA-Bootstrap-RBF-SVM	97.561	99.477	27	2.4
ANOVA-PCA-Bootstrap-L-SVM*	95.985	99.221	27	4.0
ANOVA-PCA-Bootstrap-RBF-SVM*	92.908	99.445	3	7.1
Classical SVM with linear kernel C=1	97.540	99.990	30	2.5
Classical SVM with RBF kernel C=1	97.720	99.990	30	2.3
Classical SVM with linear kernel C=0.10	97.890	99.941	30	2.1
Classical SVM with RBF kernel C=0.10	94.550	99.059	30	5.5

Source: *—Vrigazova & Ivanov, 2019, the rest is authors' calculations.

provide similar accuracy and AUC scores using all features unlike the ANOVA-BOOTSTRAP-L-SVM. The other ANOVA SVM versions (marked with asterisk) that we proposed in (Vrigazova & Ivanov, 2019) provided worse results than the ANOVA_BOOTSTRAP-L-SVM. As a result, we fitted ANOVA-BOOTSTRAP-SVM with rbf kernel that produced one of the highest accuracies in Table 3. The accuracy achieved was 97.6%, AUC score 99.5 and error rate of 2.4%. This metrics is comparable to the classical SVM with linear kernel and C=0.10. Therefore, we decided to do experiments with the ANOVA-BOOTSTRAP-RBF-SVM and find value of C that can improve the classification ability of the SVMs on the WBCD.

We first ran experiments by fitting logistic regression, decision tree, support vector machines and k-nearest neighbors using the tenfold bootstrap procedure as train/test splitting technique. As Table 4 shows the results were very similar to those in (Maldonado et al., 2014), (Khairunnahar et al., 2019; Vrigazova & Ivanov, 2019) without significant improvement of the performance metrics. We then experimented with the value of C in the rbf version of the bootstrapped ANOVA-SVM as Table 4 shows.

Table 4. Performance of modified classifications on the WBCD.

	ACC	AUC	k	Error rate
LR with bootstrap	97.362	99.494	30	2.6
DT with bootstrap	92.085	92.458	30	7.9
SVM bootstrap	97.070	99.451	30	2.9
KNN bootstrap	96.159	98.082	30	3.8
ANOVA-Bootstrap-RBF-SVM C=5	98.561	99.425	27	1.4
ANOVA-Bootstrap-RBF-SVM C=7	98.201	99.412	30	1.8
ANOVA-Bootstrap-RBF-SVM C=13	98.221	99.088	21	1.8
ANOVA-Bootstrap-RBF-SVM C=14	98.276	99.648	27	1.7
ANOVA-Bootstrap-RBF-SVM C=30	98.913	99.445	24	1.1
ANOVA-Bootstrap-RBF-SVM C=32	99.627	98.808	27	0.4

Source: authors'calculations.

Fitting ANOVA-BOOTSTRAP-RBF-SVM with C=5, 7, 13, and 14 provided accuracy from 98.2% to 98.6%. These accuracies are better than the classical classification methods and our previous attempts (Table 3). These versions of the ANOVA-BOOTSTRAP-SVM resulted in the smallest error rate compared to those in Table 3. Also, when C=13, the accuracy of 98.2% was achieved by using 21 features instead of 27 in the ANOVA-BOOTSTRAP- RBF-SVM when C=1. As Table 4 shows, the ANOVA-BOOTSTRAP-RBF-SVM with C=5, 7, 13, and 14 resulted in accuracies higher than Maldonado's (Table 1) and Khairunnahar's (Table 2). On the one hand Khairunnahar et al. (2019) used a sigmoid version of the logistic regression that did not lead to significant improved of prediction accuracy as Table 2 shows. On the other hand, as Tables 1, 3, and 4 show the support vector machines may be the most appropriate classification method on the WBDC. Therefore, the tuning of the ANOVA-BOOTSTRAP-RBF-SVM led to better performance of our proposion compared to Madonado's (Table 1).

The classical version of the support vector machines in Table 3 resulted in high accuracy (97.9%). AUC score (99.9) despite using 30 features for classification. Similar metrics result from the classical SVMs fitting with different value of C (Table 3). Maldonado's (Maldonado et al., 2014) mixed linear integer versions of the support vector classifier improve the results from the classical SVM by resulting in accuracy of 97.9 and AUC 97.7 (Table 1) by using 6 features. As he shows in his research (Maldonado et al., 2014) his versions of SVM are faster than the classical SVMs. He managed to improve the accuracy on the WDCD by reaching 98.1% and AUC score 97.9 by using 26 features (Table 1).

As Table 4 shows our proposed ANOVA-BOOTSTRAP-RBF-SVM with C=30 also outperforms Maldonado;s best accuracy of 98.1% and our previous experiments (Table 3) as well other bootstrapped classification methods (Table 4). When C=30, the accuracy reached 98.9%, while AUC score was 99.4% using only 24 features. This version of the proposed method resulted in the smallest error rate (1.1%) compared to previous work. As we show, tuning the ANOVA-BOOTSTRAP-SVM significantly outperformed classical and modified classification methods, as well as other versions of the support vector classifier.

Although error rate of 1.1% and accuracy of 98.9% are very good metrics for SVM's performance, we found a value of C, which reduced the error rate to 0.4%. We achieved this result when C=32 as Table 4 shows. The accuracy rate we achieved was 99.6% using 27 features. As Tables 1–4 show, existing research could not reach these metrics. The support vector classifier and its existing modifications were able to reach maximum of 98.1% accuracy, while we improved the classification ability of the SVMs by going beyond 99% accuracy. As Tables 1–4 show our modification outperforms not only the support vector classifier but also some other existing

classification methods and their versions. Moreover, we achieved the lowest error rate (0.4%) so far using classification methods, particularly the SVMs.

As we show in Vrigazova and Ivanov (2019), the ANOVA-BOOTSTRAP-SVM tends to be faster than Maldonado's suggestions. As Table 5 suggests the ANOVA-BOOTSTRAP-RBF-SVM needs 0.07s to produce error rate of 0.4%, while Maldonado's mixed linear integer versions of SVM (marked in green) are much slower. Much slower are also the classical versions of the SVMs (shown in yellow).

Table 5. Time comparison of SVMs's versions on the WBCD.

	Time
l2-SVM	0.20
LP-SVM	0.10
MILP1-NFS	0.20
MILP2-NFS	0.30
Fisher + SVM	0.20
l1-SVM	0.40
RFE-SVM	0.40
l0-SVM	0.50
MILP1-FS	0.20
MILP2-FS	0.20
ANOVA-CV-L-SVM	0.04
ANOVA-CV-RBF-SVM	0.05
ANOVA-Bootstrap-L-SVM	0.05
ANOVA-Bootstrap-RBF-SVM	0.07
ANOVA-PCA-Bootstrap-L-SVM	0.07
ANOVA-PCA-Bootstrap-RBF-SVM	0.06
Classical SVM with linear kernel C=1	0.65
Classical SVM with RBF kerne C=1	0.95
Classical SVM with linear kernel C=0.1	0.55
Classical SVM with RBF kernel C=0.1	1.02

Source: green—Maldonado et al., 2014; authors' calculations.

As Tables 1–5 show when tuned, the ANOVA-BOOTSTRAP-RBF-ANOVA outperformed other classification methods, provided the highest accuracy, the smallest error rate and the fastest execution time. Therefore, we managed to improve Maldonado's best result of 98.1% accuracy, 1.9% error rate and execution time of 0.20s. We contribute to academic literature by showing that the support vector machines can successfully lead to accuracy higher than 99%, reduce error rate and computational time.

We also tested our algorithm the Wisconsin Prognostic Breast Cancer dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)) and the Mammographic Mass dataset (http://archive.ics.uci.edu/ml/datasets/mammographic+mass). These datasets have a binary variable denoting benign and malignant breast cancer. We ran experiments with the ANOVA-BOOTSTRAP-RBF-SVM and compare its performance to the classical cross-validated support vector machines. Table 6 presents the results.

Table 6. Comparison of the ANOVA-BOOTSTRAP-RBF-SVM performance and the classic ANOVA-SVMs with cross validation.

Algorithm	Datset	Kernel	C	ACC	AUC	N of features	Error rate
ANOVA-BOOTSTRAP-RBF-SVM	WPBC	rbf	30	85.4	71.9	20	14.6
		rbf	5	82.3	70.3	26	17.7
		rbf	7	84.5	71.0	20	15.5
		rbf	13	83.7	71.5	26	16.3
		rbf	14	87.8	75.8	23	12.2
		rbf	32	84.6	71.9	26	15.4
	Mamographic Mass dataset	rbf	5	83.3	83.5	4	16.7
		rbf	7	81.5	83.5	4	18.5
		rbf	13	82.4	83.7	5	17.6
		rbf	14	82.5	83.7	5	17.5
		rbf	32	84.5	83.7	5	15.5
		rbf	30	81.7	83.8	3	18.3
Classic ANOVA SVMs with tenfold cross validation	WPBC	rbf	30	78.5	71.7	26	21.5
		rbf	5	74.6	69.1	3	25.4
		rbf	7	74.9	69.1	3	25.1
		rbf	13	75.1	69.1	23	24.9
		rbf	14	75.4	69.1	30	24.6
		rbf	32	78.5	71.7	26	21.5
	Mamographic Mass dataset	rbf	5	78.7	85.8	3	21.3
		rbf	7	79.0	85.9	3	21.0
		rbf	13	79.5	86.4	4	20.5
		rbf	14	79.6	86.3	4	20.4
		rbf	32	80.0	87.0	5	20.0
		rfb	30	80.1	86.9	5	19.9

Source: author's calculations.

We have used the same kernel and values of C in the classic version of the ANOVA-SVM. The only difference between our algorithm and the classic one is the mechanism to avoid overfitting. In the ANOVA-BOOTSTRAP-RBF-SVM, we use the tenfold bootstrap procedure unlike the classic ANOVA-SVM, where tenfold cross validation is used. At each fold, the bootstrap procedure uses different number and indices of the observations to perform classification. Therefore, the ANOVA-BOOTSTRAP-RBF-SVM avoids overfitting.

Table 6 shows that the cross validated ANOVA-SVM resulted in the highest accuracy of 78.5% on the Wisconsin Prognostic Breast Cancer dataset. This accuracy was achieved by setting the C parameter to 30 or 32 and leaving twenty-six features of the dataset. In comparison, the ANOVA-BOOTSTRAP-RBF-SVM achieved the highest accuracy of 87.8% on the WPBC dataset using 23 features. Moreover, the AUC score of the classic algorithm was 71.7, while the bootstrapped version produced AUC score of 75.8. Our proposed algorithm improved significantly both the accuracy (ACC) and the AUC score of the model without overfitting. Table 6 also shows that irrespective of the value chosen for C, the classic ANOVA-SVM resulted in lower accuracy and AUC scores on the WPBC dataset than the

bootstrapped ANOVA-SVM. The ANOVA-BOOTSTRAP-RBF-SVM showed better performance on the WPBC dataset than the cross-validated ANOVA-SVM in terms of the error rate. Our proposed method significantly reduced the error rate (12.2%) of the model compared to the standard method (21.5%).

Similar are the results on the Mamographic Mass dataset. The classic model produced the highest accuracy of 80.1% and AUC score of 86.9, while the bootstrapped version resulted in the best accuracy of 84.5% and AUC score equal to 83.7. Both methods used all features in the dataset. The ANOVA-BOOTSTRAP-RBF-SVM improved the accuracy on the dataset at the cost of AUC scores. However, the AUC score remained above 80, so it was high enough to consider the model good. Our model could separate correctly benign cells from malignant ones in 83.7% of the cases, while the classic one—in 86.9% of the cases. Despite this, the bootstrapped version classified correctly the class of the observation in 84.5% in the cases against 80.1% in the classic version. Table 6 shows that the bootstrapped ANOVA-SVM improved the accuracy on the Mammographic Mass dataset irrespective of the value of the parameter C. However, in all cases this came at the cost of small loss of AUC scores. The cross-validated ANOVA-SVM produced lower accuracy but higher AUC score in all cases. Despite this, the loss of AUC score in the bootstrapped version was not significant and it resulted in the smallest error rate (15.5% against 19.9% in the classic version). We consider the ANOVA-BOOTSTRAP-RBF-SVM to have a better performance on the Mammographic Mass dataset than the cross-validated ANOVA-SVM as it: 1. Produced higher accuracy score, 2. high enough AUC score and 3. lower error rate.

Tables 1–4 show that our algorithm resulted in higher accuracy scores and AUC score on the Wisconsin Diagnostic Breast Cancer dataset and lower error rate compared to other authors. An important finding from our research is that the bootstrap procedure increases the accuracy of the classification model and lowers its accuracy rate. However, it can either increase or slightly decrease the AUC score. In case, it increases the AUC score, we can accept the model as better than the classic version. In case, there is a slight decrease in the AUC score, we can still accept the model as better. If the AUC score decreases significantly, the model should be tuned, e.g. switch the kernel, the value of C, etc. or another model can be more suitable for the given dataset. Our experiments on other datasets (Vrigazova et al., 2019) show that rarely does the AUC score deteriorates significantly as a result of the bootstrap procedure although not impossible. As we show in Table 5, the ANOVA-BOOTSTRAP-RBF-SVM can also be faster than some classical models, while avoiding overfitting due to its resampling nature.

5 Conclusion and discussion

In this research we propose the ANOVA-BOOTSTRAP-RBF-SVM algorithm to increase prediction accuracy on breast cancer datasets like the Wisconsin Breast cancer dataset (99.6%) and the Mammographic Mass dataset (84.5%). With this method we aim to propose another version of the ANOVA-SVM that can improve the quality of detecting malignant breast cancer. The advantages of our proposed algorithms include improvement of the accuracy, reducing the error rate and producing high enough AUC score. Our algorithm can separate the benign from malignant cells and classify them properly with high accuracy without overfitting. As we show, in some cases our algorithm can be faster than existing ones.

It should be noted, however, that our proposition is sensitive to the kernel and the value of C in the ANOVA-SVM algorithm. Despite this, the ANOVA-BOOTSTRAP-RBF-SVM outperforms existing SVM versions. A further research can be made into the precision, recall and f1-score measures to observe the model's performance on each class as the accuracy measure is a general measure. Each class performance can be affected differently depending on the resampling procedure. Our previous experiments show that if the ability of the model to correctly classify the two binary outcomes had not improved, then the accuracy could not be improved. Therefore, we did not show the precision, recall and f1-score measures in our experiments.

With the results from this research, we extend the practical applications of the bootstrap procedure for detecting breast cancer as the ANOVA-BOOTSTRAP-RBF-SVM provided the smallest error rate, high enough AUC score and could reduce execution time in some cases. With this, we extend existing academic literature and propose optimization of the SVMs for detecting malignant and benign breast cancer that can be applied on various datasets. Moreover, the benefits from our algorithm can be extended to other medical datasets to improve ability to detect other medical conditions.

Acknowledgements

The author expresses her gratitude to prof. Ivan Ivanov for the valuable recommendations and advice.

References

Asri, H., Mousannif, H., Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064–1069.

Bashir, S., Qamar, U., & Khan, F. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Qual. Quant., 49(5), 2061–2076.

Breit, C., Ablah, E., Ward, M., Okut, H., & Tenofsky, P. (2019). Breast cancer risk assessment in patients who test negative for a hereditary cancer syndrome. The American Journal of Surgery, 219(3), 430–433.

Chaurasia, V., & Pal, S. (2007). Data mining techniques: To predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10–23.

Cortes, C., & Vapnik V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical Physics, 34(11), 4164–4172.

Forsyth, A., Barzilay, R., Hughes, K., Lui, D., Lorenz, K., Enzinger, A., Tulsky, J., & Lindvall, C. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. Journal of Pain and Symptom Management, 55(6), 1492–1499.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. Ca A Cancer Journal for Clinicians, 61(2), 69–90.

Khairunnahar, L., Hasib, M., Rezanur, R., Islam, M., & Hosain, K. (2019). Classification of malignant and benign tissue with logistic regression. Informatics in Medicine Unlocked, 16. https://doi.org/10.1016/j.imu.2019.100189.

Liu, N., Qi, E., Xu, M., Gao, B., & Liu, G. (2019). A novel intelligent classification model for breast cancer diagnosis. Information Processing & Management, 56(3), 609–623.

Mammographic Mass Dataset. Retrieved from http://archive.ics.uci.edu/ml/datasets/mammo graphic+mass.

Maldonado, S., Pérez, J., Weber, R., & Labbé, M. (2014). Feature selection for support vector machines via mixed integer linear programming. Information Sciences, 279, 163–175.

Mustafa, M., Nornazirah, A., Salih, F.M., Illzam, E., Suleiman, M., & Sharifa, A. (2016). Breast cancer: Detection markers, prognosis, and prevention. IOSR Journal of Dental and Medical ences, 15(8), 73–80.

Noske, A., Anders, S., Ettl, J., Hapfelmeier, A., Steiger, K., Specht, K., Weichert, W., Kiechle, M., & Klein, E. (2020). Risk stratification in luminal-type breast cancer: Comparison of Ki-67 with EndoPredict test results. The Breast, 49, 101–107.

Quinlan, J. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Artifitial Intelligence Research, 4(1), 77–90.

Salama, G., Abdelhalim, M., & Zeid, M. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. 1(1), 8.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence Medicine, 18(3), 205–219.

Siegel, R.L., Miller, K.D., & Jemal, A. (2015). Cancer statistics, 2015. Ca A Cancer Journal for Clinicians, 65(1), 5–29.

Singh, B. (2019). Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm. Biocybernetics and Biomedical Engineering, 39(2), 393–409.

Ting, F., Tan, Y., & Sim, K. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications, 120, 103–115.

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. Medical Hypotheses, 135. https://doi.org/10.1016/j.mehy.2019.109503.

Trieu, Ph., Tapia, K., Frazer, H., Lee, W., & Brennan, P. (2019). Improvement of cancer detection on mammograms via BREAST test sets. Academic Radiology, 26(12), e341–e347.

Vrigazova, B., & Ivanov, I. (2019). Optimization of the ANOVA procedure for support vector machines. International Journal of Recent Technology and Engineering, 8(4), 5160–5165.

Wang, H., Zheng, B., Yoon, S., & Ko, H. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, 267(2), 687–699.

Wang, P., Song, Q., Li, Y., Lv, Sh., Wang, J., Li, L., & Zhang, H. (2020). Cross-task extreme learning machine for breast cancer image classification with deep convolutional features. Biomedical Signal Processing and Control, 57. https://doi.org/10.1016/j.bspc.2019.101789.

Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2019). An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86. https://doi.org/10.1016/j.asoc.2019.105941.

Wisconsin Diagnostic Breast Cancer Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

Wisconsin Prognostic Breast Cancer Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic).

Wu, M., Zhong, X., Peng, Q., Xu, M., Huang, S., Yuan, J., Ma, J., & Tan, T. (2019). Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. European Journal of Radiology, 114, 175–184.

Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., & Zhang, F. (2019). Breast cancer histopathological image classification using a hybrid deep neural network. Methods, 1733, 52–60.

Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, T., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics, 132. https://doi.org/10.1016/j.ijmedinf.2019.103985.