computers

MDPI

Article

Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method

Ganjar Alfian [1], Muhammad Syafrudin [2,*], Imam Fahrurrozi [1], Norma Latif Fitriyani [3], Fransiskus Tatas Dwi Atmaji [4], Tri Widodo [5], Nurul Bahiyah [6], Filip Benes [7] and Jongtae Rhee [8]

[1] Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[2] Department of Artificial Intelligence, Sejong University, Seoul 05006, Korea
[3] Department of Data Science, Sejong University, Seoul 05006, Korea
[4] Industrial and System Engineering School, Telkom University, Bandung 40257, Indonesia
[5] Department of Information Technology Education, Universitas Teknologi Yogyakarta, Yogyakarta 55285, Indonesia
[6] Jurusan Ilmu Al-Qur'an dan Tafsir, IAIN Syekh Nurjati, Cirebon 45132, Indonesia
[7] Department of Economics and Control Systems, Faculty of Mining and Geology, VSB–Technical University of Ostrava, 708 00 Ostrava, Czech Republic
[8] Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, Korea
* Correspondence: udin@sejong.ac.kr; Tel.: +82-2-3408-1879

Abstract: Developing a prediction model from risk factors can provide an efficient method to recognize breast cancer. Machine learning (ML) algorithms have been applied to increase the efficiency of diagnosis at the early stage. This paper studies a support vector machine (SVM) combined with an extremely randomized trees classifier (extra-trees) to provide a diagnosis of breast cancer at the early stage based on risk factors. The extra-trees classifier was used to remove irrelevant features, while SVM was utilized to diagnose the breast cancer status. A breast cancer dataset consisting of 116 subjects was utilized by machine learning models to predict breast cancer, while the stratified 10-fold cross-validation was employed for the model evaluation. Our proposed combined SVM and extra-trees model reached the highest accuracy up to 80.23%, which was significantly better than the other ML model. The experimental results demonstrated that by applying extra-trees-based feature selection, the average ML prediction accuracy was improved by up to 7.29% as contrasted to ML without the feature selection method. Our proposed model is expected to increase the efficiency of breast cancer diagnosis based on risk factors. In addition, we presented the proposed prediction model that could be employed for web-based breast cancer prediction. The proposed model is expected to improve diagnostic decision-support systems by predicting breast cancer disease accurately.

Keywords: breast cancer; support vector machine; extra-trees; risk factors

check for updates

Citation: Alfian, G.; Syafrudin, M.; Fahrurrozi, I.; Fitriyani, N.L.; Atmaji, F.T.D.; Widodo, T.; Bahiyah, N.; Benes, F.; Rhee, J. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers* **2022**, *11*, 136. https://doi.org/10.3390/computers11090136

Academic Editor: Paolo Bellavista

Received: 9 August 2022
Accepted: 8 September 2022
Published: 12 September 2022

1. Introduction

Large-scale high-dimensional data sets have recently become accessible in a wide range of disciplines and technologies; i.e., machine learning (ML) models were employed to help analyze the medical data so that potential health issues can be identified [1–4]. One of the major global health problems and a major cause of mortality in breast cancer. The most prevalent cancer in women and one of the leading causes of mortality among them is breast cancer [5]. The World Health Organization (WHO) reported that three out of every ten women diagnosed with breast cancer globally died in 2020 [6]. Most breast cancer disease is discovered during routine screening due to its silent development [7]. The incidence, mortality, and survival rates of breast cancer could be influenced by various factors, i.e., environment, genetic factors, lifestyle, and population structure [8]. The likelihood of survival is extremely high when breast cancer is discovered and treated quickly.

Early detection of disease can be achieved by developing a prediction model so that the patient will get better treatment. Machine-learning-based models have been utilized in previous studies for detecting breast cancer and showed significant performance [9–14]. Support vector machine (SVM) is an ML model that divides instances of each class from the others by locating the linear optimum hyperplane after nonlinearly mapping the original data into a high-dimensional feature space. SVMs have demonstrated superior performance for breast cancer detection as compared to conventional models [15–18]. Additionally, earlier research has demonstrated the beneficial effects of using extra-tree as the feature selection approach to increase classification accuracy in natural language processing [19], white blood cell classification [20], and network intrusion detection [21].

However, none of these previous studies have integrated prediction models based on SVM with extra-trees into web-based breast cancer prediction. Therefore, the current study integrated SVM and extra-trees into web-based breast cancer prediction to improve the prediction performance for breast cancer. Extra-trees was utilized to extract significant risk factors, while SVM was used as a classifier to generate a more accurate prediction. In addition, integrating the proposed model into web-based breast cancer prediction could help the medical team in the decision-making process. Early prediction of breast cancer can be obtained by the medical team so that preemptive actions for the patients can be taken earlier than the incidents occur. The contributions of the present study are as follows:

(i) For the first time, we suggest a combined extra-trees and SVM technique for predicting breast cancer.
(ii) By employing extra-trees to identify the most useful features, we enhanced the performance of the proposed model.
(iii) We undertook in-depth experiments comparing the proposed model to other prediction models and findings from earlier research.
(iv) We analyzed the effects of using extra-trees or not in the feature selection approach on the accuracy performance of the model.
(v) Finally, we created a web-based breast cancer prediction tool to illustrate the viability of our model.

Additionally, the developed application can be helpful for practitioners and decision-makers as beneficial guidelines for creating and putting into practice breast cancer prediction models for practical applications.

The remainder of our study is structured as follows: ML models for breast cancer are presented in Section 2, including related SVM and extra-trees feature selection. The proposed breast cancer prediction model is described in Section 3, and Section 4 describes the experimental results and deployment of our model. Section 5 presents the conclusion including study limitations and future research directions.

2. Related Works

Machine learning (ML) models have been utilized as a prediction for disease by employing individuals' risk factors as input features. Previous studies have shown that the prediction model could improve breast cancer pre-diagnosis in many populations. Breast cancer diagnosis based on an SVM approach and feature selection was proposed by Akay [15]. The Wisconsin breast cancer dataset (WBCD) was utilized to justify the performance of their model. The experimental result showed that the proposed SVM achieved the highest classification accuracy as much as 99.51% compared to other ML models. Based on anthropometric and routine blood analysis data, Patrício et al. [16] suggested a prediction model for breast cancer. Between 2009 and 2013, they gathered 52 healthy volunteers and 64 breast cancer patients from the gynecology department of the University Hospital Center of Coimbra (CHUC). The data contain several clinical features, i.e., HOMA; levels of glucose, insulin, adiponectin, leptin, MCP-1, and resistin; and BMI. This Coimbra Breast Cancer dataset (CBCD) is publicly available and is used as a benchmark for other studies on the detection of breast cancer. ML algorithms such as random forests (RF), logistic regression (LR), and SVM were implemented as breast cancer prediction models. The results showed that SVM outperformed other models with the highest sensitivity ranging between 82% and 88%. Akben [9] proposed a decision trees model for breast cancer diagnosis on the Coimbra dataset. In their work, the decision tree used the Gini index to determine the attribute importance level. The result showed that the proposed diagnostic system has a 90.52% accuracy rate as compared to other models such as adaptive boosting (AdaBoost), SVM, K-nearest neighbor (KNN), naive Bayes (NB), artificial neural network (ANN), etc. Dalwinder et al. [10] proposed neural network with ant lion optimization to classify breast cancer. The proposed model has evaluated well-known cancer datasets, such as the Coimbra dataset. Their model utilized a wrapper method based on ant lion optimization algorithm to find the optimal feature for multilayer neural network. Their model achieved the highest accuracy 82.79% as compared to the previous studies.

Other models have been developed for breast cancer prediction by utilizing the Coimbra Breast Cancer dataset (CBCD). Zuo et al. [11] suggested curvature-based feature selection (CFS) paired with a fuzzy inference system (TSK+) as an effective filter-based feature selection technique. To predict the selected and normalized features, the proposed model was compared with other classifiers, namely KNN, AdaBoost, back-propagation neural network (BPNN), RF, linear SVM, Gaussian naïve Bayes (GNB), LR, and decision tree (DT). The result showed that their model reached the highest accuracy rate up to 85% as contrasted to other models. A voting convergent difference neural network (V-CDNN) was proposed by Zhang et al. [12] in 2021 to detect breast cancer using the Coimbra dataset. To quantify and assess the relative relevance of various traits, their study used the Gini coefficient, which is based on the random forest algorithm. Based on the experimental result, the four best features, i.e., age, glucose, body mass index (BMI), and resistin, were chosen for input variables. The proposed method was compared with traditional BPNN and achieved the highest testing accuracy as contrasted to other models.

In the machine learning area, ensemble methods have been utilized for prediction models, especially for breast cancer, and have shown positive results. The effectiveness of machine learning algorithms to predict breast cancer in women was compared by Austria et al. [13]. ML models such as LR, KNN, SVM, DT, RF, gradient boosting method (GBM), and NB were utilized as classification models for the Coimbra dataset. The result showed that GBM as one of the ensemble methods was the best classifier in predicting breast cancer with up to 74.14% accuracy rate. In addition, the result showed that BMI and levels of glucose were the top classifiers that may be an excellent pair of features for breast cancer prediction. Nanglia et al. [14] utilized an ensemble model for breast cancer prediction on the Coimbra dataset. In their study, stacking was applied for building the ensemble model consisting of three ML algorithms, namely SVM, DT, and KNN. Their proposed ensemble model reached the greatest accuracy score up to 78% as compared to other classifiers used in their study. In addition, the top five features such as levels of glucose, resistin, BMI, insulin, and homeostasis model assessment (HOMA) were obtained through the chi-square method.

Finally, previous works have shown that SVM could be adopted as a breast cancer prediction model. Rahman et al. [17] proposed computer-aided identification of breast cancer based on the ML approach. Their research used the Coimbra dataset and an SVM model with a radial basis function (RBF) kernel to identify breast cancer. The result showed that the proposed model successfully classified breast cancer by employing six features, such as BMI, age, levels of glucose, MCP-1, insulin, and resistin. Alnowami et al. [18] utilized three ML algorithms, i.e., DT, RF, and DT, and combined them with a sequential backward-selection model. The result showed that the optimal set of biomarkers such as levels of glucose, BMI, resistin, age, and HOMA can be utilized as features for the SVM model. Their proposed model achieved specificity and sensitivity rates of as much as 0.90 and 0.94, respectively.

To improve prediction performance, feature selection methods have been utilized. Previous studies showed that extra-trees have been utilized as a feature selection method to improve machine learning quality and efficiency. Nicula et al. [19] assessed the effectiveness of machine learning models for natural language processing (NLP), particularly paraphrase identification, utilizing hand-crafted features mixed with extra-trees and Siamese neural networks employing pre-trained BERT, RNNs, and BiLSTM-based models. The result showed that the extra-trees model reduced the feature space so that it obtained the best results as compared to other configurations. Baby et al. [20] utilized extra-trees as a feature selection method for Leukocyte classification. The study focused on automatic white blood cell (WBC) detection and classification on images from three different datasets. The final features were extracted based on extra-trees, and SVM was used as a classifier. The result showed that the proposed model delivers 90.76% prediction accuracy. By integrating extra-trees as feature selection and an ensemble of extreme learning machines (ELM), Sharma et al. [21] proposed an intrusion detection system. The accuracy of the developed model on the UNSW and KDDcup99 datasets, respectively, was 98.24% and 99.76%, outperforming all other techniques, according to the results.

3. Methodology

3.1. Dataset

In this study, we used the breast cancer dataset provided by previous studies [16,22]. The dataset was collected from the Gynaecology Department of the University Hospital Centre of Coimbra (CHUC) between 2009 and 2013. The Coimbra Breast Cancer Dataset (CBCD) consists of 64 women with breast cancer and 52 healthy subjects. The 9 (nine) potential risk factors (attributes) were obtained from routine blood analysis, such as body mass index/ BMI (kg/m^2), age (years), levels of glucose (mg/dL), homeostasis model assessment (HOMA), insulin (µU/mL), adiponectin (µg/mL), leptin (ng/mL), MCP-1 (pg/dL), and resistin (ng/mL). The class label (breast cancer) was assigned when the subject had positive mammography and was histologically confirmed. The purpose of this study is to diagnose whether a subject will have breast cancer in the future. Therefore, we proposed SVM and extra-trees model to predict whether a subject will later develop breast cancer.

Risk factor significance reflects the relationship between attributes and the subject's disease class. We used Pearson's correlation coefficient to investigate this relationship, which varies from -1 to +1, with a negative or positive value indicating a negative or positive correlation and a larger absolute value indicating a stronger correlation, respectively. Attributes with high correlation to the output class could be utilized as input features to maximize prediction model accuracy. Figure 1 shows that levels of glucose, insulin, HOMA, and resistin have a high positive correlation toward the class, whereas leptin has a poor correlation.

3.2. Breast Cancer Prediction Model

The proposed model consists of SVM, and extra-trees were utilized to predict breast cancer; the details can be seen in Figure 2. In our study, we employed data pre-processing to remove inappropriate and inconsistent data. During the pre-processing stage, data normalization was applied by rescaling numeric attributes into [0, 1]. The extra-trees feature selection method was utilized to remove irrelevant features, while SVM-based prediction was used as a classifier. Performance was evaluated by comparing the proposed with other machine learning models. In this study, the stratified 10-fold cross-validation (CV) was utilized for the proposed and other machine learning models. K-fold CV works by splitting the dataset into k subsets of equal size, and the instances for each subset or fold are randomly selected. Each subset, in turn, is used for testing and the remainder for the training set. The model is evaluated k times such that each subset is used once as the test set. In stratified k-fold cross-validation, each subset is stratified so that they contain approximately the same proportion of class labels as the original dataset.

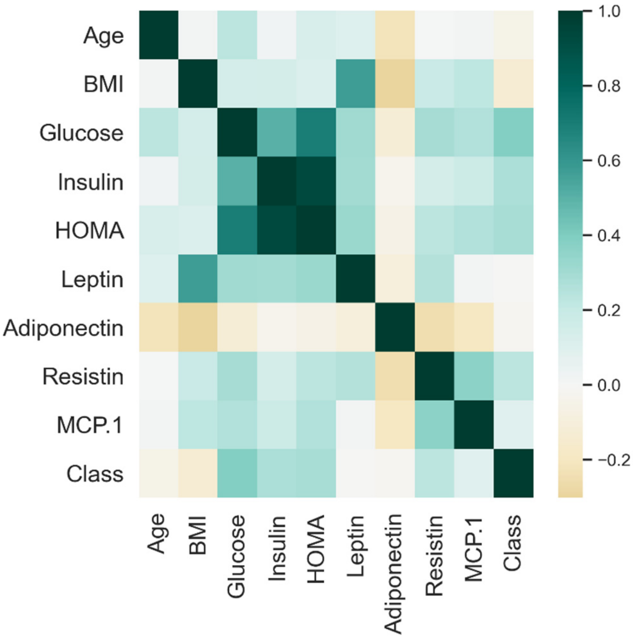


Figure 1. Attribute correlation for breast cancer dataset.

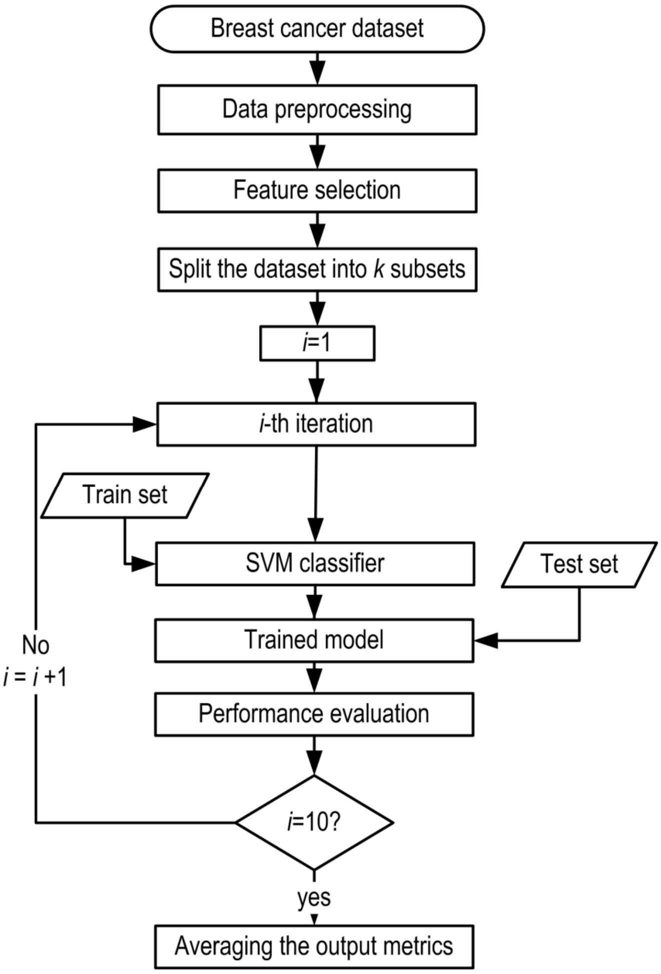


Figure 2. Flow diagram of the proposed model for breast cancer prediction.

3.3. Extra-Trees Feature Selection Method

Feature selection removes redundant and irrelevant features to improve machine learning quality and efficiency. The feature selection can be categorized into three methods: they are wrapper, filter, and embedded methods [23]. In our study, we utilized the extra-trees algorithm as one of the examples of an embedded method to extract relevant features [24].

Extra-trees generate a large number of individual decision trees from the whole training dataset. For the root node, the algorithm chooses a split rule based on a random subset of features (K) and a partially random cut point. It selects a random split to divide the parent node into two random child nodes. This process is repeated in each child node until reaching a leaf node. A leaf node is a node that does not have a child node. The predictions of all the trees are combined to establish the final prediction through a majority vote. To perform feature selection, during the construction of the forest, for each feature, the Gini importance is computed. Each feature is ordered in descending order according to the Gini importance of each feature. Finally, the user selects the top k features according to his/her choice to be used as input for the classification model.

Figure 3 shows the attribute ranking for the breast cancer dataset generated by extra-trees model. We investigated the importance of the features so that the significant attributes can be used for the classification model input. Finally, we found that the top three breast cancer features are the features that can maximize model accuracies, namely levels of glucose, age, and resistin.

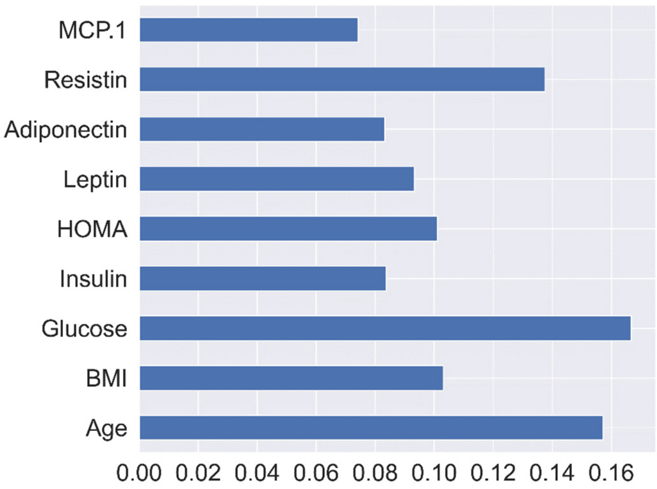


Figure 3. Feature importance using extra-trees classifier on breast cancer dataset.

3.4. SVM

In this study, the SVM algorithm was utilized as a prediction model for breast cancer. SVM can be extended to solve nonlinear classification tasks when the original data cannot be separated linearly. By applying kernel functions, the original data are mapped onto a high-dimensional feature space, in which the linear classification makes it possible to separate instances of each class from the others [25]; for the case of separating training vectors belonging to two linearly separable classes,

$$(\mathbf{x}_i\ , y_i),\ \mathbf{x}_i \in\ R^n,\ y_i\ \in \{+1, -1\},\ i = 1, \ldots, n\ , \tag{1}$$

where $\mathbf{x}_i$ is a real-valued n-dimensional input vector, and y_i is the class label associated with the training vector. The separating hyperplane is determined by an orthogonal vector $\mathbf{w}$ and bias b, which identify points that satisfy

$$\mathbf{w} \cdot \mathbf{x} + b = 0. \tag{2}$$

Thus, the classification mechanism for SVM can be expressed as

$$max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i.x_j\right)\right], \tag{3}$$

with constraints

$$\sum_{i=1}^{n}\alpha_i y_j = 0,\ 0 \le \alpha_i \le C,\ i = 1,2,\ldots,n, \tag{4}$$

where α is the parameter vector for the classifier hyperplane, and C is a penalty parameter to control the number of misclassifications [15].

We implemented the machine learning models in Python V3.7.3, utilizing the Scikit-learn V0.22.2 library [26]. In our SVM model, we set the regularization parameters $C = 1$ and radial basis function (RBF) as kernel K. The default parameters from Scikit-learn were used for other classification models. In addition, we selected the maximum number of features in the extra-trees classifier as max_features = sqrt(n_features). Therefore, the final maximum number of features is three out of nine features. The experiments were performed on an Intel Core i5-4590 computer with 8 GB RAM running Windows 7 64-bit.

Table 1 shows the confusion matrix, a useful tool to analyze classifier performance. True positive (TP) and true negative (TN) represent data that are correctly classified, whereas false positive (FP) and false negative (FN) represent data incorrectly classified. For this dataset, the patients that are diagnosed with breast cancer are labeled as 1, while normal patients are labeled as 0. Average performance metrics, such as accuracy (%), precision (%), sensitivity or recall (%), specificity (%), and area under the receiver operating characteristic curves (AUC) were obtained by conducting 10 runs of stratified 10-fold CV. Table 2 shows the classification model performance metrics based on the average value for all cross-validations.

Table 1. Classifier confusion matrix.

		Predicted Class	
		1	**0**
Actual Class	**1**	TP	FN
	0	FP	TN

Table 2. Classifier model performance metrics.

Metric	**Formula**
Accuracy	$\frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$
Precision	$\frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$
Specificity	$\frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})}$
Sensitivity/Recall	$\frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$

4. Results and Discussions

4.1. Breast Cancer Model Performances

We assessed how well the machine learning model performed and how feature selection affected the model's accuracy. The proposed SVM with extra-trees was compared with other data-driven models to predict breast cancer using known risk factors. The ML algorithms, namely logistic regression (LR), multi-layer perceptron (MLP), decision tree (DT), K-nearest neighbor (KNN), random forest (RF), naïve Bayes (NB), eXtreme Gradient Boosting (XGBoost), and adaptive boosting (AdaBoost) were employed as breast cancer prediction models. Averaging over 10 iterations of stratified 10-fold CV, the metrics for model performance are displayed in Table 3. Our proposed model combining SVM with extra-trees was higher in accuracy, precision, specificity, sensitivity, and AUC rates by up to 80.23%, 82.71%, 78.57%, 78.57%, and 0.78, respectively. Our proposed model outperformed other models for all metrics except for the recall, where XGBoost showed the better result. Finally, the proposed model achieved a 13.61% average accuracy improvement as contrasted with the other breast cancer prediction models.

Table 3. Performance metrics for breast cancer prediction.

Method	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity/Recall (%)	AUC
MLP	66.82	66.48	66.60	66.60	0.67
LR	57.58	59.90	56.14	56.14	0.56
KNN	65.91	65.73	65.52	65.52	0.66
DT	67.96	70.04	67.83	67.83	0.68
NB	57.73	62.11	59.79	59.79	0.60
RF	67.80	70.14	67.02	67.02	0.67
AdaBoost	74.09	76.51	73.43	73.43	0.73
XGBoost	75.08	75.78	74.31	82.62	0.74
Proposed model	80.23	82.71	78.57	78.57	0.78

We further evaluated the model performance by considering the receiver operating characteristic curve (ROC) as a specific metric for an unbalanced dataset [27]. False-positive and false-negative results are contrasted via the ROC curve, where AUC ≈ 1 is considered the best model [28]. Figure 4 displays the ROC curves analysis for the suggested and additional categorization models under consideration. Our findings indicated that the suggested model had the highest AUC, which was 0.78.

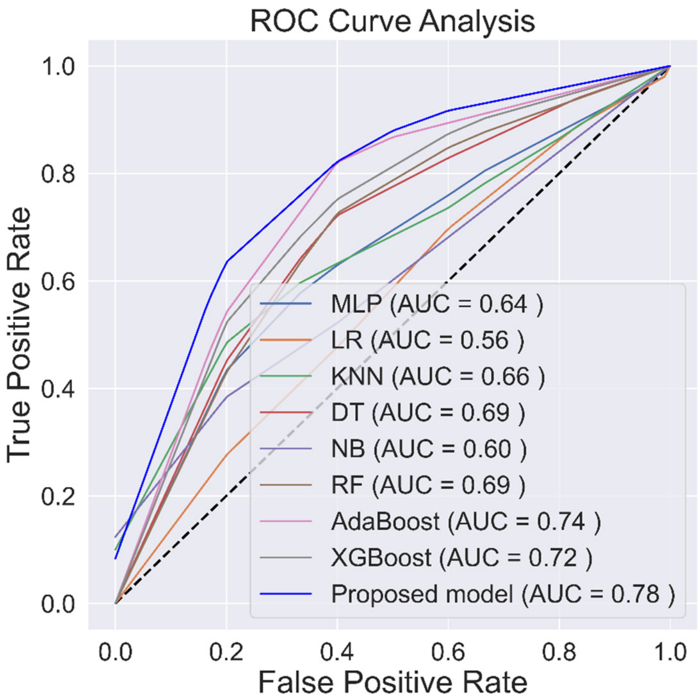


Figure 4. ROC analysis for the breast cancer prediction models.

4.2. Impact Analysis of Extra-Trees-Based Feature Selection

Figure 5 shows the effect of feature selection on the precision of classification models. The outcome showed that improved accuracy was offered by using the feature selection strategy for prediction models as opposed to using all attributes as input except for AdaBoost. Our findings demonstrated that we could increase classifier accuracy by deleting unimportant characteristics. Last but not least, adding additional tree-based feature selection increases average accuracy by up to 7.29% when compared to classifiers without the feature selection method.

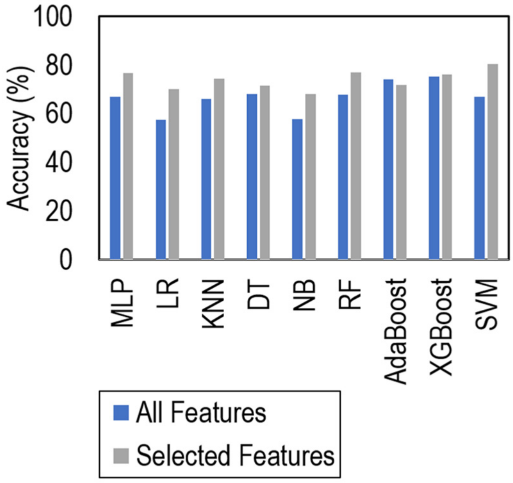


Figure 5. Impact of feature selection on classification accuracy.

Worldwide risk variables differ, and prior studies demonstrated that significant predictors for breast cancer may be extracted. This proposed study used extra-trees classifier to select the optimal features. The extra-trees algorithm identified the top three risk factors for breast cancer, which are levels of glucose, age, and resistin. These results were consistent with the previous findings to improve the diagnosis of breast cancer [16].

4.3. Study Comparison with Earlier Works

In this study, we compared the findings to those of studies that used the same breast cancer Coimbra dataset in the past. Table 4 presents a comparison of the findings between our study and earlier research.

Ghani et al. [29] applied recursive feature elimination (RFE) for feature selection and various classification models such as DT, KNN, NB, and ANN. The hold-out validation method (80% for training and 30% for testing) was employed to evaluate the performance of each ML model. The results showed that ANN performed the greatest accuracy up to 80%.

Khatun et al. [30] evaluated four ML models, such as NB, RF, MLP, and simple LR, and applied the hold-out method by splitting the data into 80% training and 20% testing. Their study revealed that MLP outperformed other ML models by achieving up to 85% accuracy rate.

Nanglia et al. [14] utilized the ensemble model and chi-square-based feature selection for breast cancer prediction on the Coimbra dataset. They formed a stacking ensemble model using three ML algorithms, such as SVM, DT, and KNN, and applied a 20-fold CV for the validation. The greatest accuracy was achieved by the model by as much as 78% as contrasted to other classifiers employed in their work.

Rasool et al. [31] developed a model with RFE for the same Coimbra breast dataset. They applied the hold-out validation (80% for training and 20% for testing) method and achieved the highest accuracy up to 76.42% for the polynomial SVM classifier.

MLP continues to have the highest classification accuracy on the Coimbra dataset [30]; however, it should be emphasized that, in contrast to 10-fold cross-validation, they used the hold-out validation approach (80%/20% for training and testing), which is less trustworthy and increases the likelihood of over-fitting and over-optimism [32]. Furthermore, none of the earlier studies offered a real-world web-based application of their research in practice. We therefore constructed and implemented a web-based application of our model for breast cancer prediction in this study.

Table 4. Comparison of our study with previous works.

Author/Method	Feature Selection	Validation Method	Accuracy (%)	Practical Application
Ghani et al. (2019) [29]/ANN	RFE	Hold-out (70/30)	80	No
Khatun et al. (2021) [30]/MLP	-	Hold-out (80/20)	85	No
Nanglia et al. (2022) [14]/Ensemble model	Chi-square	20-fold CV	78	No
Rasool et al. (2022) [31]/Polynomial SVM	RFE	Hold-out (80/20)	76.42	No
Our study/SVM and Extra-trees	Extra-trees	10-fold stratified CV	80.23	Yes

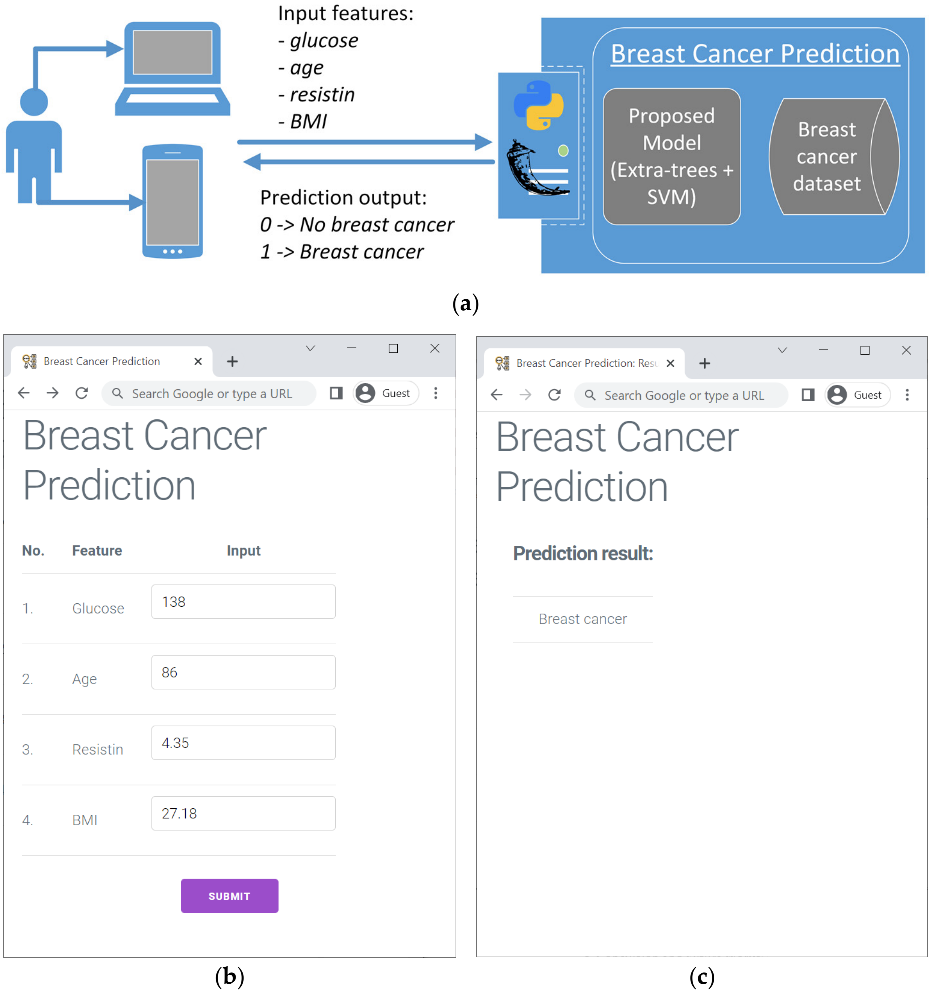
Notes: ANN, artificial neural networks; MLP, multilayer perceptron; RFE, recursive features elimination; CV, cross-validation.

It is important to note that as the reported results were obtained using various classification models, parameter settings, and validation techniques, comparing their direct performance is not fair. As a result, the results shown in Table 4 may not only be used to support the effectiveness of the categorization models but also to compare our study to earlier research in general.

4.4. Management Implications

Web-based diagnostics have been widely utilized by researchers to detect risks and facilitate decision making in a range of contexts, including the prediction of chronic disease [33,34], violent behavior [35], self-care [36], and preventive medicine [37]. Therefore, the objective of our work is to design and implement a web-based cancer screening tool that will aid the medical team in making screening decisions. The developed web-based breast cancer prediction was implemented in Python V3.9 and Flask V2.2, while the proposed prediction model was implemented using Scikit-learn V1.1.1 on the server side. Figure 6a illustrates how a user (medical team) can access an application through their web browser on a computer or a mobile device. The user can then utilize the diagnosis form as an input feature to submit it. The input feature data are then transmitted to a web server, and our proposed model is employed to diagnose the subjects' breast cancer status. The diagnosis result is then presented to the user in the prediction output interface. The proposed breast cancer prediction model is developed from breast cancer data and implements an extra-tree model for feature selection and SVM to predict the final breast cancer status.

The diagnosis form interface that users can input is shown in Figure 6b. When all the required fields have been filled out, the user can click the "submit" button to send the information to a secure remote server, which loads our model to predict the subjects' risk of breast cancer. The resulting interface, as depicted in Figure 6c, receives the prediction status after that. This application is expected to help individuals with an early breast cancer diagnosis and improve the performance of breast cancer classification. Therefore, preventive actions or further treatments can be provided to each individual.



Breast Cancer Prediction

No.	Feature	Input
1.	Glucose	138
2.	Age	86
3.	Resistin	4.35
4.	BMI	27.18

Breast Cancer Prediction

Prediction result:

Breast cancer

Figure 6. The developed web-based breast cancer prediction: (**a**) designed framework; (**b**) input form interface; (**c**) prediction output interface.

5. Conclusions and Future Works

Our study proposed a breast cancer prediction model based on SVM and extra-trees. A dataset incorporating breast cancer risk factors was employed. Our proposed combined SVM and Extra-trees model was contrasted with other ML prediction models, i.e., LR, NB, KNN, DT, RF, AdaBoost, MLP, and XGBoost. Our study showed that the proposed model outperformed other models, achieving accuracy = 80.23%. The extra-trees classifier was used to identify significant features from the dataset. Furthermore, by utilizing extra-trees as a feature selection method, the average ML prediction accuracy was improved by up to 7.29% as contrasted to ML without a feature selection method. In addition, we integrated our prediction model into web-based breast cancer prediction. This web-based system can be utilized to support the medical team in the decision-making practice regarding breast cancer. Finally, our study is expected to improve healthcare systems and help reduce the breast cancer risk for individuals.

Our study focused on a small set of the population; thus, the result may not be generalized for wider cases. A future study should consider other clinical datasets, prediction models, and feature selection methods.

Author Contributions: Conceptualization, G.A. and M.S.; methodology, G.A.; software, M.S. and N.L.F.; validation, F.T.D.A., T.W. and N.B.; formal analysis, M.S.; investigation, N.L.F.; resources, F.B. and F.T.D.A.; data curation, T.W. and N.B.; writing—original draft preparation, G.A.; writing—review and editing, G.A. and M.S.; visualization, I.F. and T.W.; supervision, J.R.; project administration, I.F.; funding acquisition, F.B. and J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset can be downloaded here: [22].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Abbreviation	Definition
SVM	Support vector machine
Extra-trees	Extremely randomized trees classifier
CV	Cross-validation
AI	Artificial intelligence
ML	Machine learning
WHO	World Health Organization
WBCD	Wisconsin Breast Cancer dataset
CHUC	University Hospital Centre of Coimbra
LR	Logistic regression
CBCD	Coimbra Breast Cancer dataset
GNB	Gaussian naïve Bayes
BPNN	Back-propagation neural network
TP	True positive
KNN	K-nearest neighbor
V-CDNN	Voting convergent difference neural network
MLP	Multi-layer perceptron
CFS	Curvature-based feature selection
RF	Random forest
DT	Decision tree
NB	Naïve Bayes
TN	True negative
GBM	Gradient boosting method
BMI	Body mass index
ANN	Artificial neural network
RBF	Radial basis function
NLP	Natural language processing
WBC	White blood cell
ELM	Extreme learning machine
AUC	Area under the receiver operating characteristic curves
FN	False negative
HOMA	Homeostasis model assessment
FP	False positive
AdaBoost	Adaptive boosting
XGBoost	eXtreme Gradient Boosting
ROC	Receiver operating characteristic curve
RFE	Recursive features elimination

References

1. Alfian, G.; Syafrudin, M.; Fitriyani, N.L.; Anshari, M.; Stasa, P.; Svub, J.; Rhee, J. Deep Neural Network for Predicting Diabetic Retinopathy from Risk Factors. *Mathematics* **2020**, *8*, 1620. [CrossRef]
2. Alfian, G.; Syafrudin, M.; Fitriyani, N.L.; Syaekhoni, M.A.; Rhee, J. Utilizing IoT-Based Sensors and Prediction Model for Health-Care Monitoring System. In *Artificial Intelligence and Big Data Analytics for Smart Healthcare*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 63–80. ISBN 978-0-12-822060-3.
3. Fitriyani, N.L.; Syafrudin, M.; Alfian, G.; Rhee, J. Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. *IEEE Access* **2019**, *7*, 144777–144789. [CrossRef]

4. Fitriyani, N.L.; Syafrudin, M.; Alfian, G.; Fatwanto, A.; Qolbiyani, S.L.; Rhee, J. Prediction Model for Type 2 Diabetes Using Stacked Ensemble Classifiers. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 399–402.
5. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef] [PubMed]
6. Breast Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 15 August 2021).
7. Alkabban, F.M.; Ferguson, T. Breast Cancer. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
8. Hortobagyi, G.N.; de la Garza Salazar, J.; Pritchard, K.; Amadori, D.; Haidinger, R.; Hudis, C.A.; Khaled, H.; Liu, M.-C.; Martin, M.; Namer, M.; et al. The Global Breast Cancer Burden: Variations in Epidemiology and Survival. *Clin. Breast Cancer* **2005**, *6*, 391–401. [CrossRef] [PubMed]
9. Akben, S. Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System. *IRBM* **2019**, *40*, 355–360. [CrossRef]
10. Dalwinder, S.; Birmohan, S.; Manpreet, K. Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer. *Biocybern. Biomed. Eng.* **2019**, *40*, 337–351. [CrossRef]
11. Zuo, Z.; Li, J.; Xu, H.; Al Moubayed, N. Curvature-based feature selection with application in classifying electronic health records. *Technol. Forecast. Soc. Chang.* **2021**, *173*, 121127. [CrossRef]
12. Zhang, Z.; Chen, B.; Xu, S.; Chen, G.; Xie, J. A novel voting convergent difference neural network for diagnosing breast cancer. *Neurocomputing* **2021**, *437*, 339–350. [CrossRef]
13. Austria, Y.D.; Lalata, J.-A.; Maria, L.B.S., Jr.; Goh, J.E.; Goh, M.L.; Vicente, H. Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. *Int. J. Simul. Syst. Sci. Technol.* **2019**, *20*, 23.1–23.8. [CrossRef]
14. Nanglia, S.; Ahmad, M.; Khan, F.A.; Jhanjhi, N. An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomed. Signal Process. Control* **2021**, *72*, 103279. [CrossRef]
15. Akay, M.F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **2009**, *36*, 3240–3247. [CrossRef]
16. Patrício, M.; Pereira, J.; Crisóstomo, J.; Matafome, P.; Gomes, M.; Seiça, R.; Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer* **2018**, *18*, 29. [CrossRef]
17. Rahman, M.; Ghasemi, Y.; Suley, E.; Zhou, Y.; Wang, S.; Rogers, J. Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features. *IRBM* **2020**, *42*, 215–226. [CrossRef]
18. Alnowami, M.R.; Abolaban, F.A.; Taha, E. A Wrapper-Based Feature Selection Approach to Investigate Potential Biomarkers for Early Detection of Breast Cancer. *J. Radiat. Res. Appl. Sci.* **2022**, *15*, 104–110. [CrossRef]
19. Nicula, B.; Dascalu, M.; Newton, N.N.; Orcutt, E.; McNamara, D.S. Automated Paraphrase Quality Assessment Using Language Models and Transfer Learning. *Computers* **2021**, *10*, 166. [CrossRef]
20. Baby, D.; Devaraj, S.J.; Hemanth, J.; M, A.R.M. Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach. *Turk. J. Electr. Eng. Comput. Sci.* **2021**, *29*, 2742–2757. [CrossRef]
21. Sharma, J.; Giri, C.; Granmo, O.-C.; Goodwin, M.; Sharma, J.; Giri, C.; Granmo, O.-C.; Goodwin, M. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP J. Inf. Secur.* **2019**, *2019*, 15. [CrossRef]
22. Breast Cancer Dataset. Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra (accessed on 1 June 2022).
23. Guyon, I. *Feature Extraction Foundations and Applications*; Springer: Berlin, Germany, 2006; Volume 207, ISBN 9783540354871.
24. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
25. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
26. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
27. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
28. Huang, J.; Ling, C. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310. [CrossRef]
29. Ghani, M.U.; Alam, T.M.; Jaskani, F.H. Comparison of Classification Models for Early Prediction of Breast Cancer. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2019; pp. 1–6.
30. Khatun, T.; Utsho, M.M.R.; Islam, M.A.; Zohura, M.F.; Hossen, M.S.; Rimi, R.A.; Anni, S.J. Performance Analysis of Breast Cancer: A Machine Learning Approach. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 1426–1434.
31. Rasool, A.; Bunterngchit, C.; Tiejian, L.; Islam, R.; Qu, Q.; Jiang, Q. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3211. [CrossRef] [PubMed]
32. Santos, M.S.; Soares, J.P.; Abreu, P.H.; Araujo, H.; Santos, J. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 59–76. [CrossRef]
33. Alfian, G.; Syafrudin, M.; Ijaz, M.F.; Syaekhoni, M.A.; Fitriyani, N.L.; Rhee, J. A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing. *Sensors* **2018**, *18*, 2183. [CrossRef] [PubMed]

34. Fitriyani, N.L.; Syafrudin, M.; Alfian, G.; Rhee, J. HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access* **2020**, *8*, 133034–133050. [CrossRef]
35. Krebs, J.; Negatsch, V.; Berg, C.; Aigner, A.; Opitz-Welke, A.; Seidel, P.; Konrad, N.; Voulgaris, A. Applicability of two violence risk assessment tools in a psychiatric prison hospital population. *Behav. Sci. Law* **2020**, *38*, 471–481. [CrossRef]
36. Syafrudin, M.; Alfian, G.; Fitriyani, N.L.; Anshari, M.; Hadibarata, T.; Fatwanto, A.; Rhee, J. A Self-Care Prediction Model for Children with Disability Based on Genetic Algorithm and Extreme Gradient Boosting. *Mathematics* **2020**, *8*, 1590. [CrossRef]
37. Yu, C.-S.; Lin, Y.-J.; Lin, C.-H.; Lin, S.-Y.; Wu, J.L.; Chang, S.-S. Development of an Online Health Care Assessment for Preventive Medicine: A Machine Learning Approach. *J. Med. Internet Res.* **2020**, *22*, e18585. [CrossRef]