Healthcare Analytics 2 (2022) 100010

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

Meta-Health Stack: A new approach for breast cancer prediction

Mina Samieinasab [a], S. Ahmad Torabzadeh [a], Arman Behnam [b], Amir Aghsami [a,c,*], Fariborz Jolai [a]

[a] School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
[c] School of Industrial Engineering, K. N. Toosi University of Technology (KNTU), Tehran, Iran

ARTICLE INFO

Keywords:
Breast cancer detection
Machine learning
Prediction
Ensemble methods
Analytics

ABSTRACT

Data analytics and machine learning have grown in importance to efficiently manage large amounts of healthcare data. Recent statistics indicate that breast cancer is the most commonly diagnosed cancer worldwide. Different tumor features are available in various datasets for breast cancer detection. Filtering those to obtain an accurate diagnosis is time-consuming and challenging. Machine learning algorithms are beneficial for finding a significant relationship between various features and malignant tumors. This research proposes a new ensemble-based framework named Meta-Health Stack to predict breast cancer more efficiently. In this framework, to extract the most relevant features, the Extra Trees classifier is used to integrate the attributes obtained from Variance Inflation Factor, Pearson's Correlation, and Information Gain to detect the tumors' hidden patterns. Finally, three approaches, including Boosting, Bagging, and Voting, were combined with equal weights together through the Stacking approach. The proposed method resulted in a 97% F1-score and 98% precision tested on Wisconsin Diagnosed Diagnostic Breast Cancer (WDBC) dataset. Based on the findings, we noticed that the suggested framework's performance works perfectly due to the selection of more appropriate features by the Extra Trees algorithm. Furthermore, we recommend that this proposed framework be used to diagnose breast cancer in its early stages as it works effectively. Using this framework, breast cancer recovery and therapy will be more successful. Moreover, to evaluate the performance of the proposed framework, it has been implemented on three other medical datasets. Results show an appropriate performance in predicting other illnesses as well.

1. Introduction and literature review

Diagnosis and prognosis of breast cancer are highly prioritized due to the importance and prevalence of this cancer type, especially in women. For example, in the United States, it accounts for about 30% of all cancer types in womenSiegel, Miller [1]. According to the statistics published by the World Health Organization (WHO) in 2018, More than two million new breast cancer cases have been identified. This number is expected to reach 2.8 million cases by 2040 - a significant statistic rise. Late treatment of breast cancer leads to harmful stages of cancer and thus lower survival rates. Thus, early detection of cancer could notably decrease its mortality rate [2].

The existence of powerful computers has facilitated the acquisition and use of data in today's world. Using data to extract useful information can lead to earlier detection, and it can also improve the power of doctors' decision-making [3]. Data mining (DM) is the process of extracting valuable knowledge from large databases. In recent years, machine learning (ML) and DM have become widespread for discovering patterns from different datasets. Moreover, they are used to develop expert systems to help physicians improve diagnosis accuracy [4]. For instance, Brause [5] conducted a study to show machine learning(ML) will enhance diagnosis accuracy. The results indicate that the most experienced physician diagnoses correctly in 79.97% of cases, while ML diagnoses with an accuracy of 91.1%.

The accurate prediction of breast cancer is among the most critical tasks for physicians because it leads to quick responses and better survival chances. Moreover, mammography requires human and material resources, which make it a complex process. Therefore, the prediction of breast cancer can be simplified with DM and ML techniques. Several expert systems have been developed for breast cancer diagnosis using different methodologies such as statistical approaches, support vector machine(SVM), neural network, fuzzy systems, and hybrid methods [6]. For breast cancer predictions, major ML algorithms such as logistic regression(LR) [7], artificial neural networks(ANN) [8], K Nearest

* Corresponding author at: School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Tehran, Iran.
E-mail addresses: mina.samieinasab@ut.ac.ir (M. Samieinasab), ahmad.torabzadeh@ut.ac.ir (S.A. Torabzadeh), arman_behnam@ind.iust.ac.ir (A. Behnam), a.aghsami@ut.ac.ir (A. Aghsami), fjolai@ut.ac.ir (F. Jolai).

https://doi.org/10.1016/j.health.2021.100010
Received 20 October 2021; Received in revised form 20 November 2021; Accepted 23 November 2021

Neighbors (KNN) [9], Decision tree(DT) [10], Random forest(RF) [11], Naïve Bayes [12], SVM [13,14], etc. are used.

Feature engineering is one of the most critical parts to improve ML methods' performance, improves our understanding of the data, reduces the model's complexity and execution time, and increases the model's performance. The feature selection is selecting the most relevant data features [15,16]. In other words, feature engineering prevents undesired correlation in the learning process by eliminating redundant and irrelevant features [17]. For example, Memon, Li [18] proposed a method that achieved higher accuracy than other state-of-the-art methods due to the implementation of appropriate feature selection. Zheng, Yoon [13] paid particular attention to feature extraction and selection for a high-quality breast cancer diagnosis classifier. They developed a hybrid of Support Vector Machine and K-means(K-SVM) algorithms. Said, Abd-Elmegid [19] used a technique to divide the dataset into different clusters based on their feature similarity and applied the classification model on these clusters rather than the full dataset. Kumara, Sushila [20] used the Heatmap matrix to show the correlation between features. The features having co-efficient values close to one must be eliminated. SVM is implemented on the extracted features as well as all features. Pasha and Mohamed [21] introduced a novel Bio-inspired Ensemble Feature Selection (BEFS) model that worked with ML and DM algorithms while relevant and essential features were selected by an ensemble algorithm named 'random forest' and a bio-inspired algorithm called genetic algorithm. Panda, Swagatika [12] used principal component analysis (PCA) to reduce the Wisconsin breast cancer dataset's dimension from nine features to four to maintain the most uncorrelated data. Ed-daoudy and Maalmi [22] used Association Rules (AR) to reduce feature dimension, and in this way, they selected eight inputs, which provided a high accuracy.

Recent advances in ML have led to the development of new methods that combine several single models and simultaneously take advantage of them. These methods usually produce more accurate solutions than a single model would. These learning methods have been called "meta-learning schemes" or "meta-classifiers" or "ensembles" [23]. Some of the popular techniques in constructing ensemble models are listed below:

- Decision Trees (DTs) were widely used to build ensemble classification models. Many types of them were implemented in different research studies, such as Simple Classification and Regression Trees (CART), C4.5, CART, Reduced Error Pruning Tree (REPTree), and Decision Stump [24]. DTs are comprehensible, they stand outliers, and they prevent over-fitting by pruning.
- The SVM algorithms are widely adopted in constructing ensembles. They have several advantages, such as avoiding over-fitting. Furthermore, they could be used for high-dimensional data, but their output is hard to interpret [25].
- ANN is mostly used in complex datasets because it has an appropriate function applying to noisy data [26].
- Another algorithm in constructing ensembles is Bayesian classifiers. It has the ability to deal with irrelevant features and missing values [27].

In the following lines, some of the articles that combined these single techniques to construct ensemble models are discussed:

Abdar and Makarenkov [28] used the confidence-weighted voting method, an ensemble classifier, which integrates an SVM with Boosting ANNs to diagnose breast cancer. After performing the ensemble method by backup vector ML algorithms and artificial neural networks, the level of accuracy obtained by CVW-BANN got higher. Basunia, Pervin [29] used a stacking classifier, an ensemble method that combines several classification techniques, classified tumors into two categories — benign and malignant. They applied classification techniques such as CART, RF, LR, KNN, and SVM; Then computed their accuracy. Kumar, Gangal [30] used bootstrap aggregation for creating sample datasets. This algorithm is an ensemble technique that reduces the variance of ML methods using the decision tree for processing data that divides the dataset into two branches. On each branch, Deep Learning (DL) and SVM algorithms are applied separately. They showed that the proposed model, based on SVM or DL combined with the decision tree, performs better than other standard bagging models. Srimani and Koti [31] conducted ensemble methods on five medical datasets and reached an excellent enhancement in the base classifiers' performance. More recently, Islam, Haque [32] compared SVM, RF, LR, KNN, and ANN according to different measures to show that ANN outperforms other algorithms in all the performance metrics.

Moreover, researchers used different tools to construct and evaluate their ensemble models. Some of these tools are open-sources such as Weka [11], R [33], Python packages [34], BVLC Caffe [35], and other types of tools are commercial such as MATLAB [36], IBM SPSS Modeler [37], and SAS Enterprise Miner [38].

Furthermore, the parameter setting of ML models has a significant impact on their performances. Therefore, it is essential to tune the ensemble classifiers' parameters. Some researchers adopted evolutionary algorithms, including Bayesian optimization and the genetic algorithm [39]. Moreover, some other algorithms were utilized for this purpose. For example, Salma [40] proposed the Bat algorithm to optimize Extreme Machine Learning (EML) parameters to achieve high accuracy. Said, Abd-Elmegid [19] used the Hyper Parameter Optimization technique to enhance the classification model's accuracy rate.

Besides Ensemble classifiers, some papers used Deep Learning techniques to predict breast cancer. These algorithms are used to process complicated and high-dimensional datasets. For instance, for intelligent image analysis, deep neural networks are widely used [41]. However, the application of the DL method is rarely seen in papers on the Wisconsin dataset. Mekha and Teeyasuksaet [42] used DL to predict breast cancer and compare the results with other classification methods such as SVM, Decision tree, Naïve Bayes (NB), Vote (DT+NB+SVM), RF, and AdaBoost. The results proved the successful performance of DL. Panda, Swagatika [12] used Deep Forest, which has few parameters and can be deployed as a fully automatic model. They overcame the problem of traditional neural networks' high number of parameters. More recently, Gupta and Garg [43] used different machine learning algorithms to classify breast cancer tumors. Additionally, they used the deep learning approach and found out that using Adam Gradient Descent Learning has the highest accuracy among all algorithms.

Based on reviewing recent literature, it appears to be some research gaps in breast cancer prediction studies. Firstly, previous studies did not pay much attention to feature engineering, and it may lead to higher training time and increase the chance of overfitting. Second, most of them used statistical methods, neither in the structure of their model nor in the performance evaluation.

This paper proposed a novel framework that has an appropriate function in predicting breast cancer. Moreover, this framework has been tested on three other datasets to show its adequate performance. More precisely, this framework selects the most appropriate features using a statistical approach to describe information better and keep our prediction model more effective. To do this, different feature selection algorithms are aggregated, and the results are compared together. Finally, we used Extra Trees to determine the best features to utilize in the classification section. In the following, the extracted features are used for breast cancer prediction. Ensemble models such as Bagging, Boosting, and Voting were combined by a mathematical stacked-based model, which weighs these three approaches equally. Moreover, instead of using accuracy, recall and F1-score were used in both feature engineering and classification sections due to the nature of our problem.

Taken together, this study highlights some new points in this area that can be useful for developing a better prediction model and helping healthcare systems to diagnose breast cancer more accurately. We introduced a new approach that leads to an effective breast cancer diagnosis and reduces the risk of missing valuable information. Besides, it could be beneficial for other types of healthcare datasets aiming to detect whether someone has an illness.

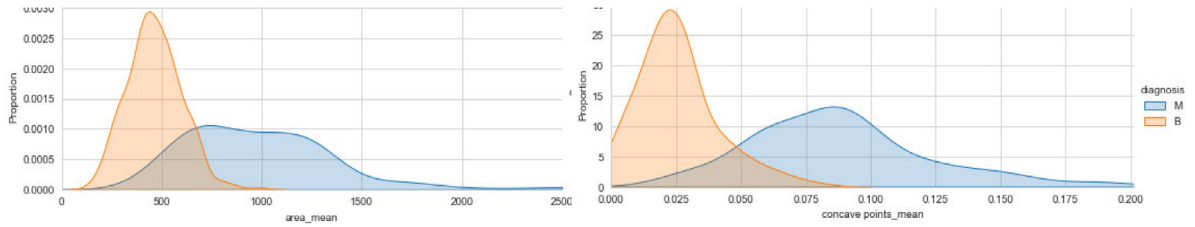

Fig. 1. (a) Area vs. Diagnosis (b) Concave points vs Diagnosis (Blue = Malignant; Orange = Benign).. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Data description.

Attributes	Description
Radius	Mean of distances from the center to points on the perimeter
Texture	The standard deviation of gray-scale values
Perimeter	The outer perimeter of the tumor in the image
Area	Tumor image covered with content
Smoothness	Local variation in radius lengths
Compactness	Perimeter^2/area – 1.0
Concavity	The severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	Image two sides symmetry
Fractal dimension	Coastline approximation'' – 1

The rest of this paper is organized as follows. Section 2 describes the dataset used in the experiment and the proposed methods, and Section 3 presents the framework's results. In Section 4, the presented framework is implemented on three other datasets to evaluate its performance. In Section 5, our findings' usefulness, consistency with other papers, limitations, and possible future works are presented. Finally, in Section 6, we concluded the paper.

2. Methodology

In this study, we designed a new framework for cancer prediction named Meta-Health Stack. This method is sensitive to Precision and F1-score. We used Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the University of California — Irvine repository [44] to show our proposed framework's performance and accuracy.

2.1. Dataset description

In this study, Wisconsin Diagnostic Breast Cancer (WDBC) dataset is used. It has 570 observations, including 30 attributes in 10 groups for each cell nucleus presented in Table 1. Features are extracted from a digital image of a breast mass. Three indicators are measured for each group: mean value, standard error, and maximum value. The three measurements for each group are considered features in the dataset. Thus, the dataset has 30 attributes. Our goal is to name label attribute, whether it is a benign tumor or a malignant one.

According to clinical literature for tumor type prediction, two vital features are area and concave points. From Fig. 1(a), tumor type is very interpretable based on Area and Concave points. When these values increase, the probability of being malignant becomes higher, while in lower values of these attributes, a benign tumor is more probable.

For more features' explanations, probability density for all features according to the target variable is shown in Fig. 2. According to this figure, the more benign and malignant cases are separate, the more the corresponding variable depends on the target variable. These variables, including concave points_mean, radius_worst, perimeter_worst, area_worst, and concave points_worst, are very interpretable. On the other hand, some variables, including compactness_se, concavity_se, concave points_se, and concavity_mean, are not interpretable due to their inseparability in their histograms. This issue is critical for any cancer prediction and should be considered in our feature engineering method. Consequently, high correlated features with the target masking other features' impact and features with less correlation with the target value are critical in our calculations and should be considered.

2.2. Feature engineering

We used both ML algorithms and metric-based methods for the feature engineering section. In medical cases, it is crucial to predict the malignant instances correctly. Features importance is a measure of how useful attributes are in predicting a target variable. Using features with the highest feature importance can provide insight into the dataset and improve a predictive model. It also helps to estimate our goal, which is to identify the malignant cases accurately.

We applied two different approaches to our dataset for feature selection. Firstly, a Heatmap correlation matrix was used to remove the most correlated features. Then, Univariate feature selection, Recursive feature elimination with (and without) cross-validation, and Tree-based feature selection were applied to the remained features separately. Secondly, Variance inflation factors (VIF), Information Gain, and Pearson's Correlation were applied to the whole dataset. Then, the Extra Trees classifier was used to aggregate the mentioned algorithms' results. We used Random Forest to see how well each of these methods works to predict our problem.

Fig. 3 shows a heatmap matrix, also known as a correlation matrix, among all features. The correlation coefficient ranges from −1 to 1. The value closer to one indicates that the features are closely correlated, and the conclusion is that the features are positively dependent on each other. In contrast, the value closer to zero indicates that the features are independent of each other. As a result, there is a perfect correlation.

2.3. Classification algorithms

The extracted features from Extra Trees are used in the classification section in Bagging, Boosting, and Voting algorithms. Three types of bagging algorithms named Bagged Classifier, Random Forest, and Extra Trees were applied. We used AdaBoost, XGBoost, Gradient Boosting, and LightGBM (LGBM) Classifier for Boosting algorithm. The AdaBoost is the first algorithm from the boosting's family, which is easy to understand and has a few hyper-parameters to be tuned. However, when the dataset has irrelevant features, Adaboost's performance is not appropriate [45]. XGBoost and Gradient Boosting Machines (GBMs) are ensemble tree-based methods that apply the principle of gradient descent. However, XGBoost has more parameters to be optimized and may enhance the prediction in this way [46]. The LGBM concept is based on a gradient boosting algorithm, but it is faster than XGBoost because of inspecting the most informative samples [47]. In the end, we aggregated the algorithms' results with optimum weights using the Stacking method. The overall procedure of our proposed method is presented in Fig. 3. We defined the concept of these methods below.

2.3.1. Bagging

A bagging algorithm is an approach in which based classifiers are trained in parallel, and each training sample is selected randomly from the whole dataset. This algorithm fits some classifiers and aggregates their prediction result by voting or averaging. Bootstrap samples are used to fit almost all independent models because it needs a large

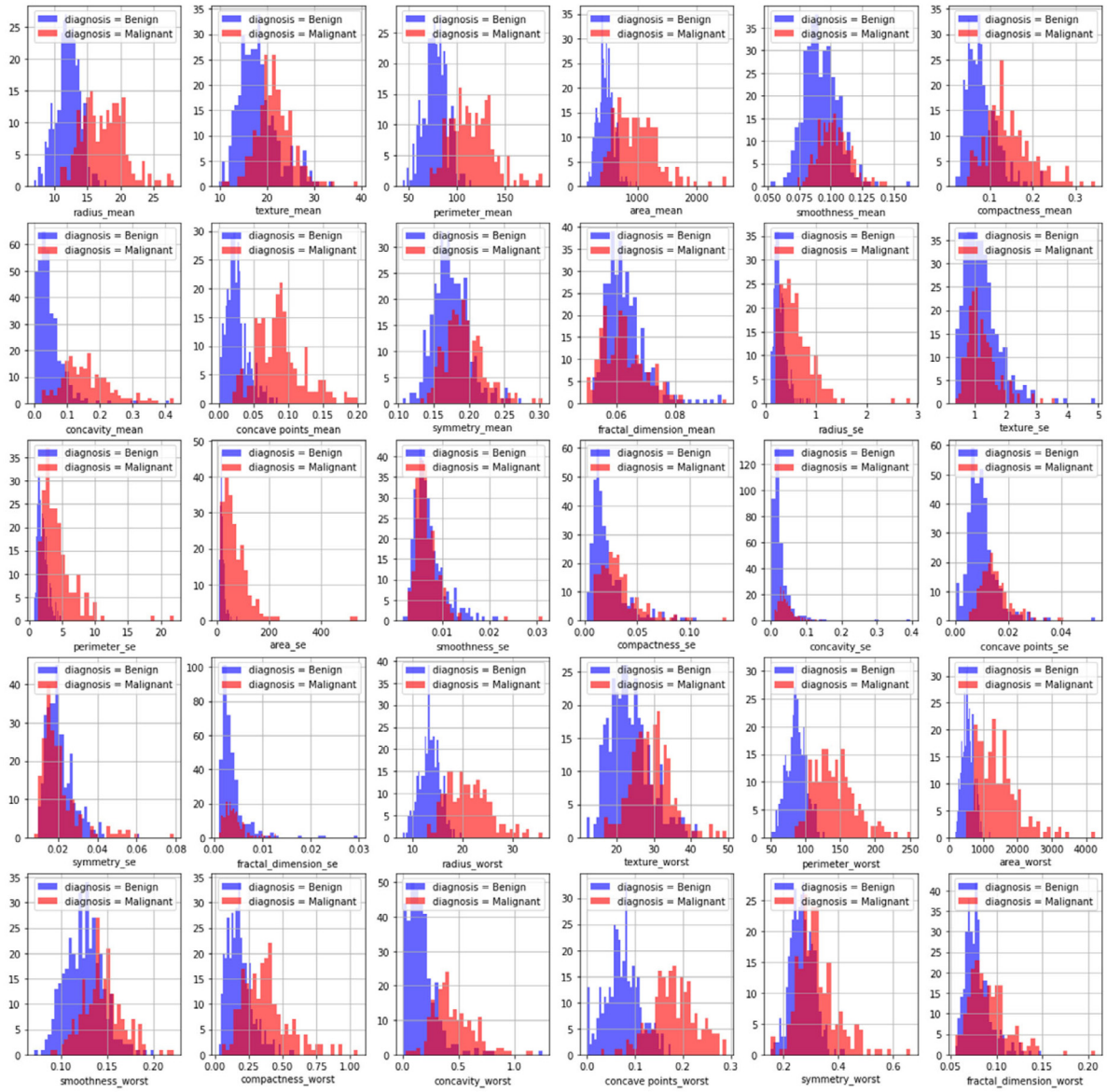

Fig. 2. The probability density for each feature according to the target variable.

amount of data to fit completely independent models. Bagging helps to build a stable learning algorithm that prevents overfitting and reduces variance. Random Forest, Extra Trees, and Bagged decision trees are among the most common bagging algorithms that are also used in this study [48].

2.3.2. Boosting

Boosting is an ensemble model that is used to reduce variance. It was introduced by Schapire 1990 [49] to boost the performance of weak learning algorithms by building strong learners from several weak ones. In the first step, this model starts with training data, and then in the second step, it tries to make the previous model better and correct the errors. This process is continued. Therefore, in Boosting, models are trained sequentially, and each model considers the success of the previous model. In this way, the model could focus on the most difficult data samples by giving higher weight or importance. Some of the most common boosting algorithms used in this study are AdaBoost, Gradient Boosting, XGboost, and LGBM Classifier.

2.3.3. Stacking

The stacking approach combines weak algorithms by training a meta-model to build a model with higher prediction accuracy. Thus, two things have to be defined in the stacking method: the classification algorithms we want to fit and the meta-model that combines them [50].

2.4. Proposed method

Our proposed method is a stacking-based model using each algorithm according to its abilities. We used two approaches in the feature engineering section of the framework to rely on the most interpretable features. Then, three ensemble approaches are stacked to create a breakthrough for predicting breast cancer. Fig. 4 presents an overall procedure of our framework.

2.5. Metrics for evaluation of performance

Accuracy, precision, recall, and F1-score as presented in Eqs. (1)–(4) are used to evaluate the different ensemble algorithms' performance with selected features. Accuracy could be used as a defining metric in many models, although, in many cases, precision and recall are advisable to be considered. In some situations, the accuracy is very high, while the precision or recall is low. But in our model, we have to prevent misclassifying a malignant tumor as a benign one. As a result,

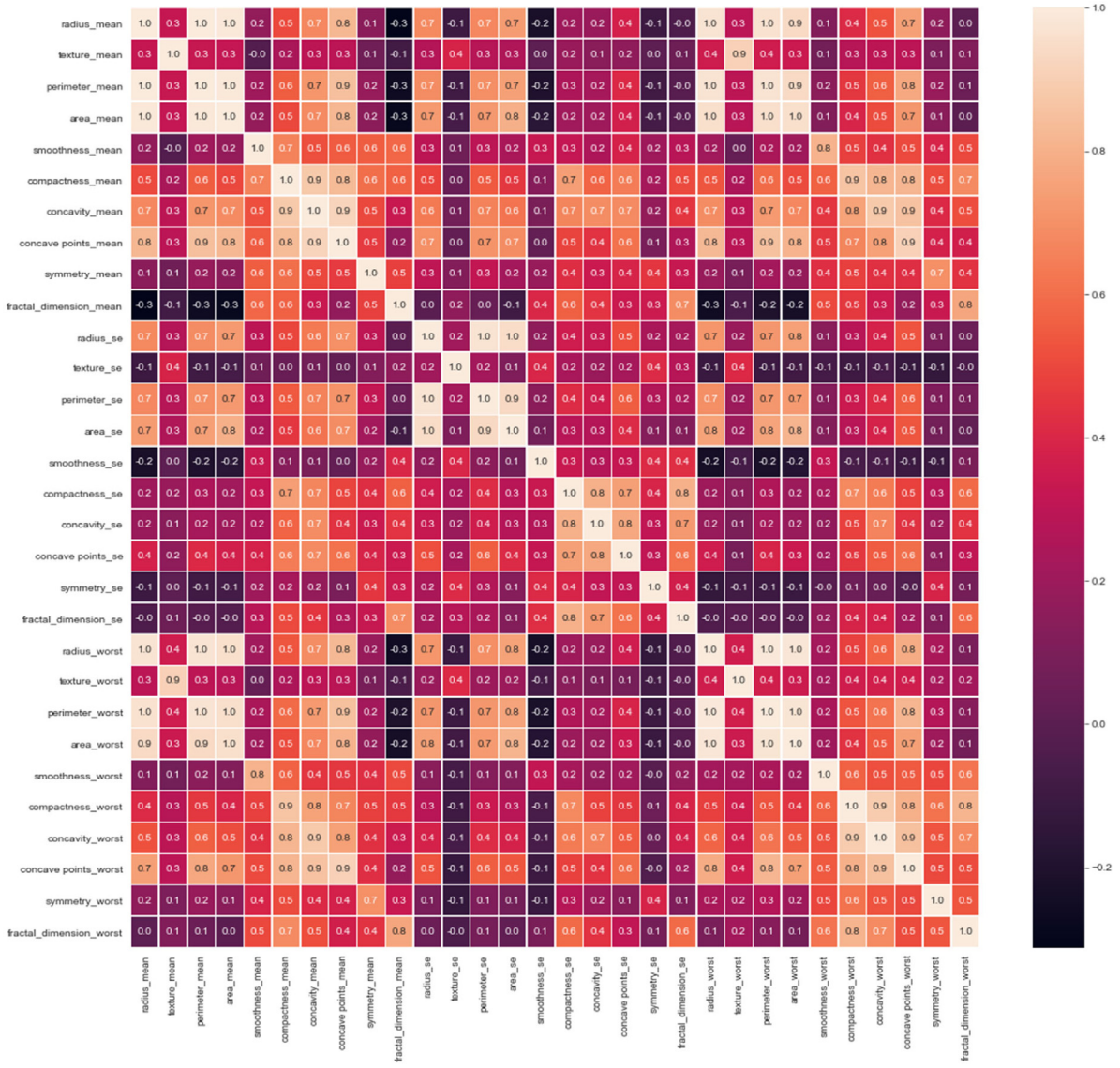

Fig. 3. Heatmap matrix of all features.

high recall is a better metric in our case. The measure of correctly recognizing true positives is called recall. Precision tells us how many true positives are identified among all positives. In some situations, precision and recall are equally important. For example, if a patient is misdiagnosed with breast cancer, treatment for his or her actual disease may be postponed. Hence, our goal is to reach a high precision as well as a high recall.

To illustrate this issue, consider a classifier that uses a variety of patient measures to predict whether a patient has a malignant tumor or a benign one. We are more concerned with this classifier's capacity to discover everyone with true malignant tumur – not letting anyone slip through the cracks, the threat to their health unnoticed – than with overall model accuracy. We may be lenient with the model's tendency to overpredict sickness because such folks will likely seek additional testing, visit their physicians, and so on, and the issue will be resolved. However, if the model predicts a benign tumor in a patient who has actually a malignant one, that is a much more serious mistake. That patient is sent home without being treated. That is why Recall is more important in medical cases. However, it is not enough for a good prediction; Because if the model predicts all the instances as malignant, then the Recall goes to 1. This issue would not be desirable. Therefore, F1score is also important and should be considered.

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{1}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$F1Score = 2 * \frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recall}} \tag{4}$$

TP and TN are the numbers of true positives and true negatives that are correctly labeled. FP and FN refer to wrongly labeled samples. The Confusion Matrix (CM) metrics present these four mentioned metrics in a table [46].

3. Results

3.1. Feature engineering results

Python, a popular open-source programming language, is used in this experiment. As discussed in Section 2, two approaches were applied for feature selection. We used the Heatmap correlation matrix in the first approach and chose 16 features to use in Univariate [51], RFE [52], RFE with cross-validation, and tree-based algorithms. In the second approach, VIF [53], Pearson's correlation [54], and Information Gain [55] were applied on all 30 features, and in the following, the Extra Trees ensemble algorithm was applied.

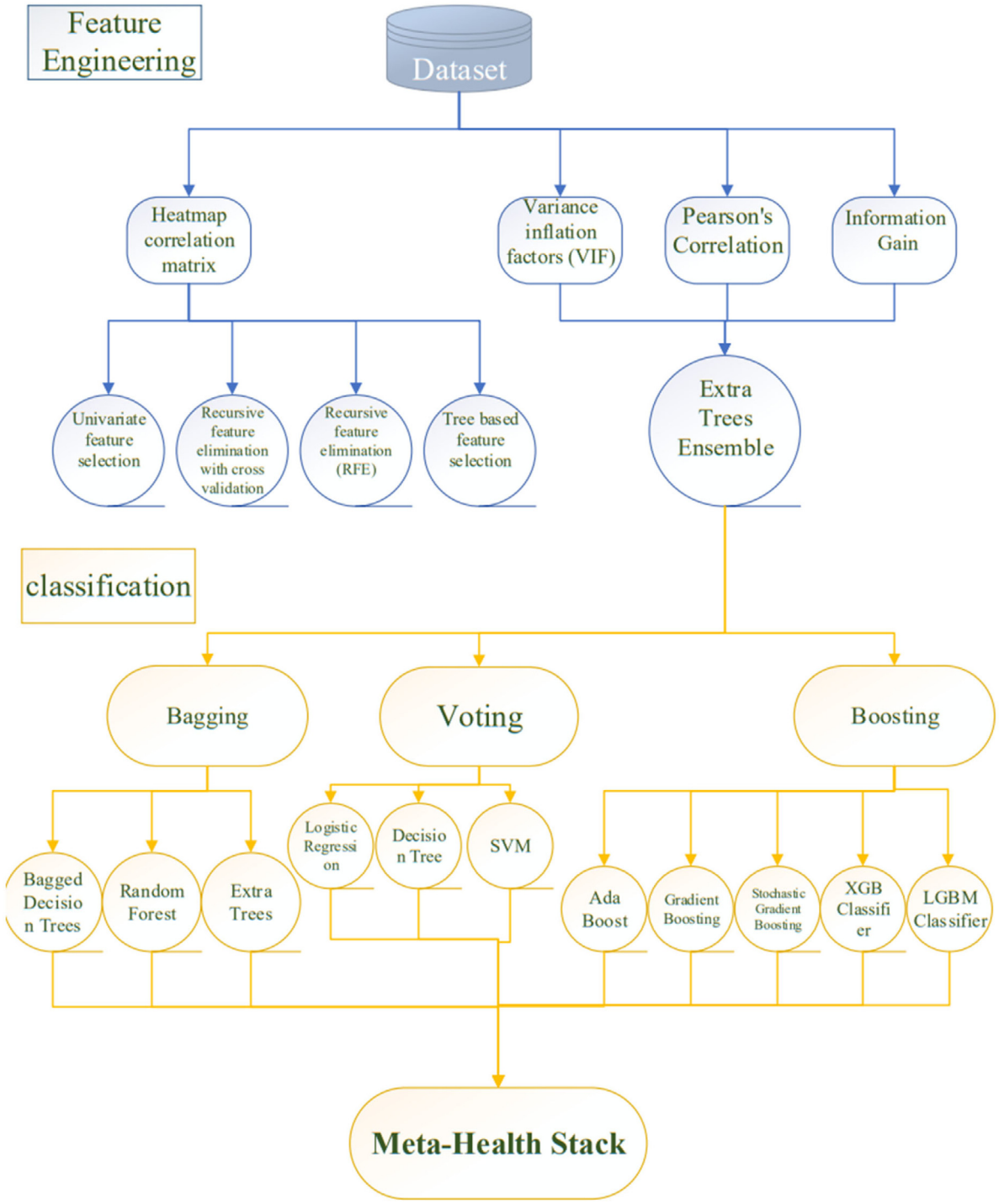

Fig. 4. The overall procedure of the proposed method.

One of the notable individual methods is implementing a tree-based algorithm on the dataset without preprocessing and using the Heatmap to remove highly correlated features. According to Fig. 5, concave points_worst, area_worst, perimeter_worst, and radius_worst have a high deviation of feature importance value in the prediction procedure; Especially the first two mentioned features. Therefore, these features are not reliable in prediction due to their high deviation in their feature importance values.

The experimental results achieved in the feature selection step are given in Table 2. All the results in Table 2 are achieved after applying Random Forest. Since recall and F1-score are the most preferred performance evaluation measure, we chose the Extra Trees algorithm to select features.

Each method gives us a threshold for features' impact and specifies the number of most important features. As we can see in Table 2, many mentioned methods such as VIF, Pearson's Correlation, Univariate, and Information Gain cannot introduce a high number of vital features. We call these methods group one. Although VIF has the highest recall and F1-score, it is useless because of the limited features it introduces. Those methods with a high number of selected features have low values in performance metrics, which we call group two. We built an Extra Tree-based on group one, which achieved high-performance metrics and introduced many important features.

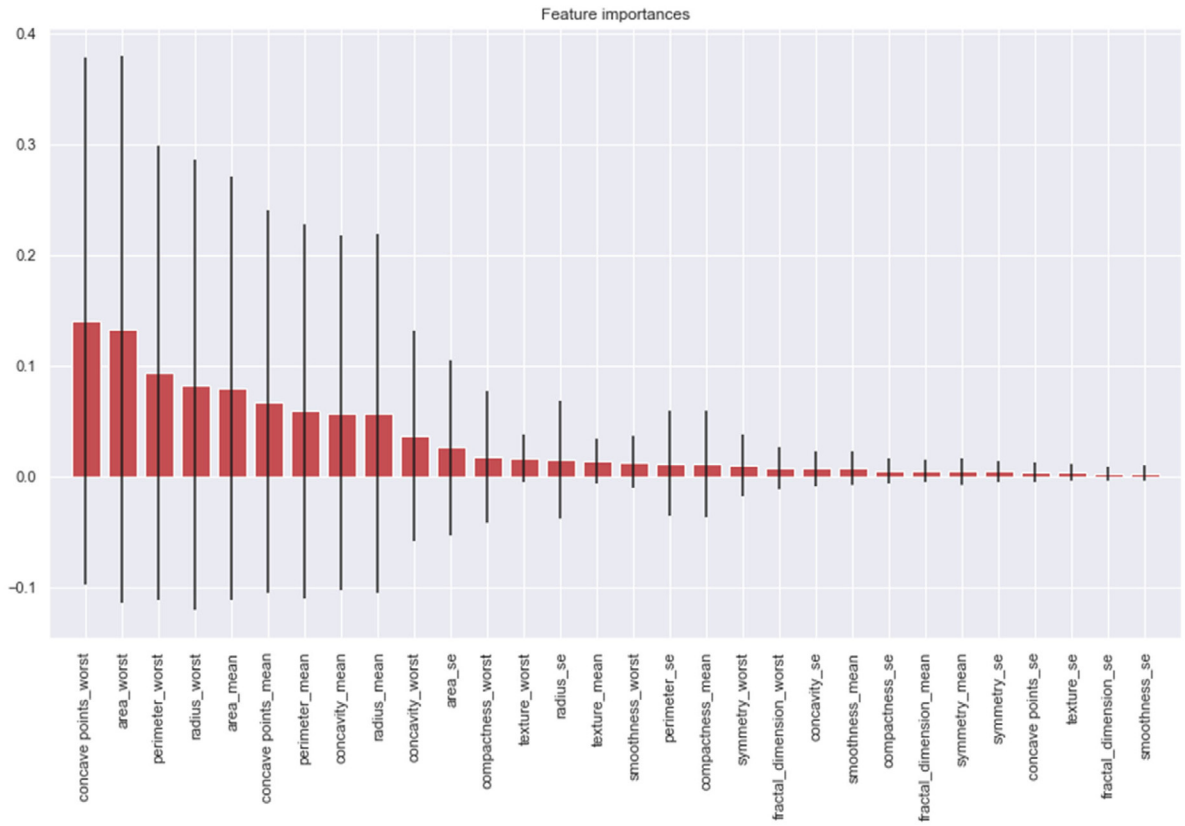

Fig. 5. Deviation in feature importance based on Random Forest.

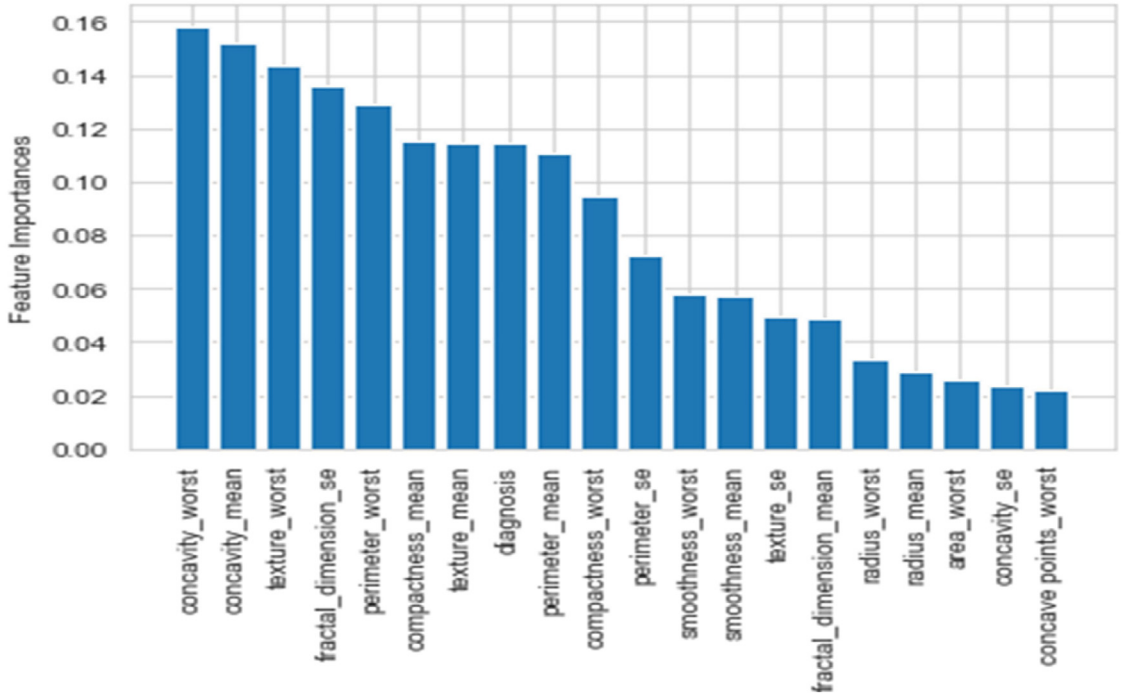

Fig. 6. Feature importance based on Extra Trees Ensemble feature selection and random forest classification.

According to Fig. 6, the feature importance values of fourteen features are higher than 0.05. As we have discussed, these features were selected to fit prediction models, and the rest of them were removed.

3.2. Prediction results

In this section, our described framework for the prediction part is implemented. Each of these algorithms' fitting performance depends on

Table 2
Performance's evaluation metrics for feature selection algorithms.

Feature selection algorithms	Precision	Recall	F1-score	Accuracy	Number of selected features
Correlation	0.96	0.93	0.95	0.96	16
Univariate	0.96	0.93	0.95	0.94	5
RFE	0.98	0.93	0.95	0.97	10
RFE + cross validation	0.96	0.93	0.95	0.96	15
Tree-based	0.92	0.92	0.92	0.94	16
VIF	0.96	1	0.98	0.99	8
Pearson's correlation	1	0.9	0.94	0.96	9
Information gain	1	0.92	0.96	0.97	6
Extra trees ensemble	**1**	**0.93**	**0.96**	**0.97**	**14**

Table 3
Classification algorithms' parameters.

Classification algorithms	Parameters
Bagged classifier	n_estimators=1000
Random forest	n_estimators=1000
Extra trees	n_estimators=1000, max_features=7
AdaBoost	n_estimators=10
GradientBoosting	n_estimators=3000, learning_rate=0.05 max_depth=4, max_features='sqrt',min_samples_leaf=15,min_samples_split=10
XGBClassifier	colsample_bytree=0.4603, gamma=0.0468, learning_rate=0.05, max_depth=3, min_child_weight=1.7817, n_estimators=2200, reg_alpha=0.4640, reg_lambda=0.8571
LGBMClassifier	objective = 'binary',boosting_type='gbdt',num_leaves=5, learning_rate=0.05, n_estimators=720, max_bin = 55, bagging_fraction = 0.8, bagging_freq = 5, feature_fraction = 0.2319, feature_fraction_seed=9, bagging_seed=9, min_data_in_leaf =6, min_sum_hessian_in_leaf = 11
Voting ensemble	LogisticRegression(solver='liblinear'), DecisionTreeClassifier, SVC (gamma='scale')

Table 4
Classification algorithms' performance evaluation metrics for test data.

Classification algorithms	Precision	Recall	F1-score	Accuracy
Bagged classifier	0.972	0.936	0.951	0.964
Random forest	0.980	0.936	0.959	0.970
Extra trees	0.984	0.968	0.976	0.982
AdaBoost	0.90	0.952	0.930	0.947
Gradient boosting	0.980	0.952	0.960	0.970
XGB classifier	0.970	0.936	0.959	0.982
LGBM classifier	0.964	0.936	0.959	0.964
Voting ensemble	0.925	0.793	0.854	0.900
Meta-Health stack	**0.985**	**0.968**	**0.976**	**0.982**

Table 5
Data description of the heart.

Attributes	Description
Age	
Sex	Sex (1 = male; 0 = female)
cp	Chest pain type (typical angina) – Value 1: typical angina – Value 2: atypical angina – Value 3: non-angina pain – Value 4: asymptomatic
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholesterol in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	Resting electrocardiographic results – Value 0: Normal – Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Coastline approximation'' – 1
ca	Number of major vessels (0–3) colored by fluoroscopy
thal	Displays the thalassemia 0 = normal; 1 = fixed defect; 2 = reversible defect

the parameters that have been chosen for them. These parameters are described below:

- n_estimators: This parameter determines the number of trees in a forest
- max_features: This parameter specifies the number of features to do the best split.
- learning_rate: This parameter specifies each tree's effect on the final result.
- max_depth: This parameter prevents over-fitting by determining the maximum depth of the tree.

As a result, the discussed parameters in Table 3 are selected as the best parameters of each classification algorithm according to a trial and error method. The algorithms that do not have specific parameters are not mentioned here.

Numerical results of applying classification algorithms are given in Table 4. As discussed in Section 2.5, F1-score and Recall play a more critical role in medical cases, and our selection for the best predictor was based on these two-evaluation metrics. According to Table 4, we used bagging and boosting approaches and their algorithms individually and compared their results. The voting approach was based on simple classifiers such as Logistic Regression, Decision Tree classifier, and Support Vector Classifier, which have different logics compared to decision trees or Bagging and Boosting systems. Finally, we stacked all the results of the predictions and fitness with the same weight. Interpreting the error in data is an essential topic that is applicable here by using the mentioned approaches with the same weights. In the case of being more confident about a specific method, the given weight to that method could be higher in the stacking approach.

Confusion matrix of feature engineering and classification results for all the methods are presented in Appendix A. All outputs, including TP, TN, FP, FN, accuracy, precision, and recall can be obtained from these figures.

Appendix B contains a great visual comparison between the methods used in this study based on accuracy, recall (the most important metric), and F1-score. This comparison is performed for test data. In these graphs, our method is compared to all Boosting, Bagging, and Voting approaches.

As seen in the first figure that compares approaches by accuracy, most of the methods do well for train data except for Voting and AdaBoost. However, the differences are about 1 to 2 percent between the accuracy values. In the second figure, all the methods perform exceptionally well in recall metrics for train data, but just Extra Trees and our method do better than others for the unseen data known as test data. The results are the same for the F1-score. Therefore, our proposed framework outperforms all other methods and approaches. For a few feature numbers, Extra Trees using Random Forest have equal performances to our approach. Nonetheless, our model is more resistant to changes in the feature number. Moreover, it has an appropriate function on other datasets to predict cancer occurrence or any other diseases.

4. Sensitivity analysis

To assess the function of our proposed framework, we implemented it on three other health-related data sets named heart, proc_heart_cleve, and prostate cancer. Similar to our main dataset, each of the mentioned datasets have some numerical and Boolean attributes, and the goal is to detect the presence of the disease. The attributes description of these datasets are given in Tables 5–7.

It should be mentioned that sex, thalach, and ca were eliminated in the heart dataset after the feature engineering process. Similarly,

Table 6
Data description of heart_proc_clev.

Attributes	Description
Age	
Sex	Sex (1 = male; 0 = female)
ind_typ_angina	Typical angina (0/1)
ind_atyp_angina	Atypical angina (0/1)
ind_non_ang_pain	*Non-angina chest pain (0/1)*
resting_BP	Rest blood pressure
Serum_cholest	*Serum cholesterol in mg/dl*
blood_sugar_exc120	*Fasting blood glucose > 120* mg/dl *(0/1)*
ind_for_ecg_1	Ecg 1 indicator (0/1)
ind_for_ecg_2	*Ecg 2 indicator (0/1)*
Max_heart_rate	*Maximum heart rate*
ind_exerc_angina	*Exercise-induced angina indicator (0/1)*
ST_dep_by_exerc	Exercise-induced ST depression with respect to rest
ind_for_slope_up_exerc	Peak exercise ST segment upward slope indicator (0/1)
ind_for_slope_down_exerc	Downslope indicator of peak exercise ST segment (0/1)
num_vessels_fluro	The number of great vessels colored by fluoroscopy (0–3)
Thal_rev_defect	Thal reversible damage indicator (0/1)
Thal_fixed_defect	Thal permanent damage indicator (0/1)

Table 7
Data description of prostate cancer.

Attributes	Description
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	The outer perimeter of the tumor in the image
Area	Tumor image covered with content
Smoothness	Local variation in radius lengths
Compactness	Perimeter^2/area – 1.0
Symmetry	Image two sides symmetry
Fractal dimension	Coastline approximation" – 1

Table 8
Classification algorithms' performance evaluation metrics for Heart data.

Classification algorithms	Precision	Recall	F1-score	Accuracy
Bagged classifier	0.801	0.804	0.801	0.802
Random forest	0.829	0.780	0.804	0.791
Extra trees	0.822	0.740	0.778	0.769
AdaBoost	0.872	0.820	0.845	0.835
Gradient boosting	0.750	0.720	0.734	0.714
XGB classifier	0.822	0.740	0.778	0.769
LGBM classifier	0.872	0.820	0.845	0.835
Voting ensemble	0.844	0.825	0.829	0.835
Meta-Health stack	0.836	0.820	0.828	0.813

Table 9
Classification algorithms' performance evaluation metrics for proc_heart_cleve data.

Classification algorithms	Precision	Recall	F1-score	Accuracy
Bagged classifier	0.742	0.739	0.742	0.755
Random forest	0.771	0.702	0.722	0.777
Extra trees	0.718	0.621	0.718	0.744
AdaBoost	0.756	0.756	0.756	0.800
Gradient boosting	0.659	0.783	0.716	0.744
XGB classifier	0.736	0.756	0.736	0.788
LGBM classifier	0.736	0.756	0.736	0.788
Voting ensemble	0.782	0.779	0.781	0.788
Meta-Health stack	0.750	0.810	0.779	0.811

max_heart_rate, ind_for_slope_up_exerc, and num_vessels_fluro in proc_heart_cleve dataset and texture, perimeter, and Smoothness from prostate cancer were removed.

After applying our proposed framework to the mentioned datasets, the classification algorithms' performance evaluation metrics for datasets are given in Tables 8–10.

5. Discussion

The proposed framework, which consists of a combination of feature engineering and classification, will lead to an effective and accurate breast cancer diagnosis. By implementing this framework by practitioners, breast cancer mortality could be reduced. Our research was limited to a dataset that had ten relevant features. Nonetheless, some more significant factors in predicting breast cancer can be extracted from other kinds of datasets. (e.g., mammographic data). Our proposed method provides higher prediction quality in comparison with similar previous studies mentioned in the literature review. For example, Zheng et al. [16] obtained 97.38% accuracy, which is lower than our proposed method's accuracy. Furthermore, in their study, accuracy and CPU time were the only considered evaluation metrics. As we mentioned before, in medical cases, recall and F1-score are more critical and must be considered carefully.

Table 10
Classification algorithms' performance evaluation metrics for Prostate Cancer data.

Classification algorithms	Precision	Recall	F1-score	Accuracy
Bagged classifier	0.750	0.784	0.761	0.800
Random forest	0.744	0.744	0.744	0.800
Extra trees	0.904	0.863	0.883	0.833
AdaBoost	0.900	0.818	0.857	0.800
Gradient boosting	0.869	0.909	0.888	0.833
XGB classifier	0.863	0.863	0.863	0.800
LGBM classifier	0	0	0	0.266
Voting ensemble	0.900	0.818	0.857	0.800
Meta-Health stack	0.863	0.863	0.863	0.800

The performance of an appropriate ensemble method depends on the selected features for the model's input and classification method selection. Our proposed model selects attributes with high feature importance as input for our prediction models. Choosing the attributes with the highest feature importance provides a global insight into the model's behavior. Moreover, all interactions with other features are automatically taken into account when calculating feature importance measures.

We chose the integrating approach to aggregate the result of some most effective ensemble methods, such as different types of bagging, boosting, and voting. Based on the results, the proposed approach to predicting breast cancer provided the best recall and F1-score compared to the previous studies. It benefits from a new method for feature reduction leading to a more precise prediction.

6. Conclusion and future works

Cancers' rapid and accurate diagnosis is one of the main challenges in medical studies. Breast cancer is the main cause of death for women worldwide. This study was conducted to improve the previous studies' classification methods in predicting breast cancer. To this aim, a framework called Meta-Health Stack was introduced. This framework consists of two parts: feature selection and classification. In the first part, the Extra Trees algorithm was used to integrate the results of VIF, Information Gain, and Pearson's Correlation methods to select the appropriate attributes as input to the classification section. In the next section, the results of Bagging, Boosting, and Voting algorithms were integrated with equal shares using the Stacking approach. The final results indicate that the proposed framework allows us to get a 97% F1-score and recall and 98% accuracy. The findings on WBCD's breast cancer dataset demonstrated that the Meta-Health Stack framework would improve the diagnosis' performance. Moreover, the framework is tested on three other medical-based datasets, and the results showed the high ability of the framework in predicting illnesses.

In the future, implementing this framework on bigger datasets and evaluating it on a larger scale, if possible, could be a positive challenge for future studies. Furthermore, The Meta-Health Stack can integrate with some optimization techniques such as GA (Genetic algorithm), PSO (particle swarm optimization), or ACO (ant colony optimization algorithm). These techniques can be utilized to choose the best parameters of ensemble algorithms precisely.

Moreover, a graphical user interface could be designed. It can provide more comfortable use of this framework by medical practitioners. The practitioners can enter all the relevant patient's data and obtain the classification result without any ML and data science experience.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their sincere thanks to the editor-in-chief and the anonymous reviewers for their valuable comments and suggestions which have led to a significant improvement of the manuscript. This work was supported by the Iranian National Science Foundation (91049559).

Appendix A

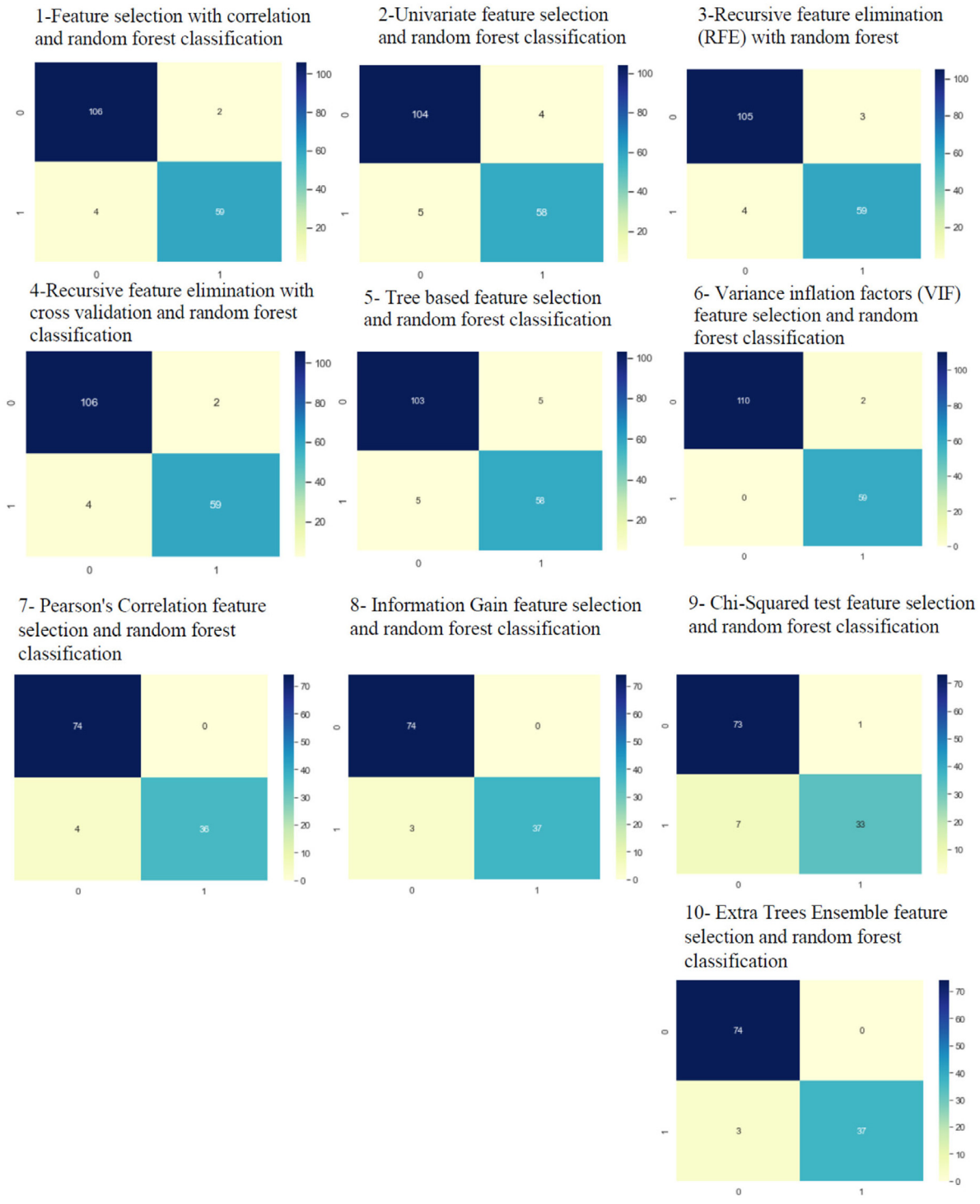

Fig. A.1. Each feature engineering method classification confusion matrix using Random Forest method.

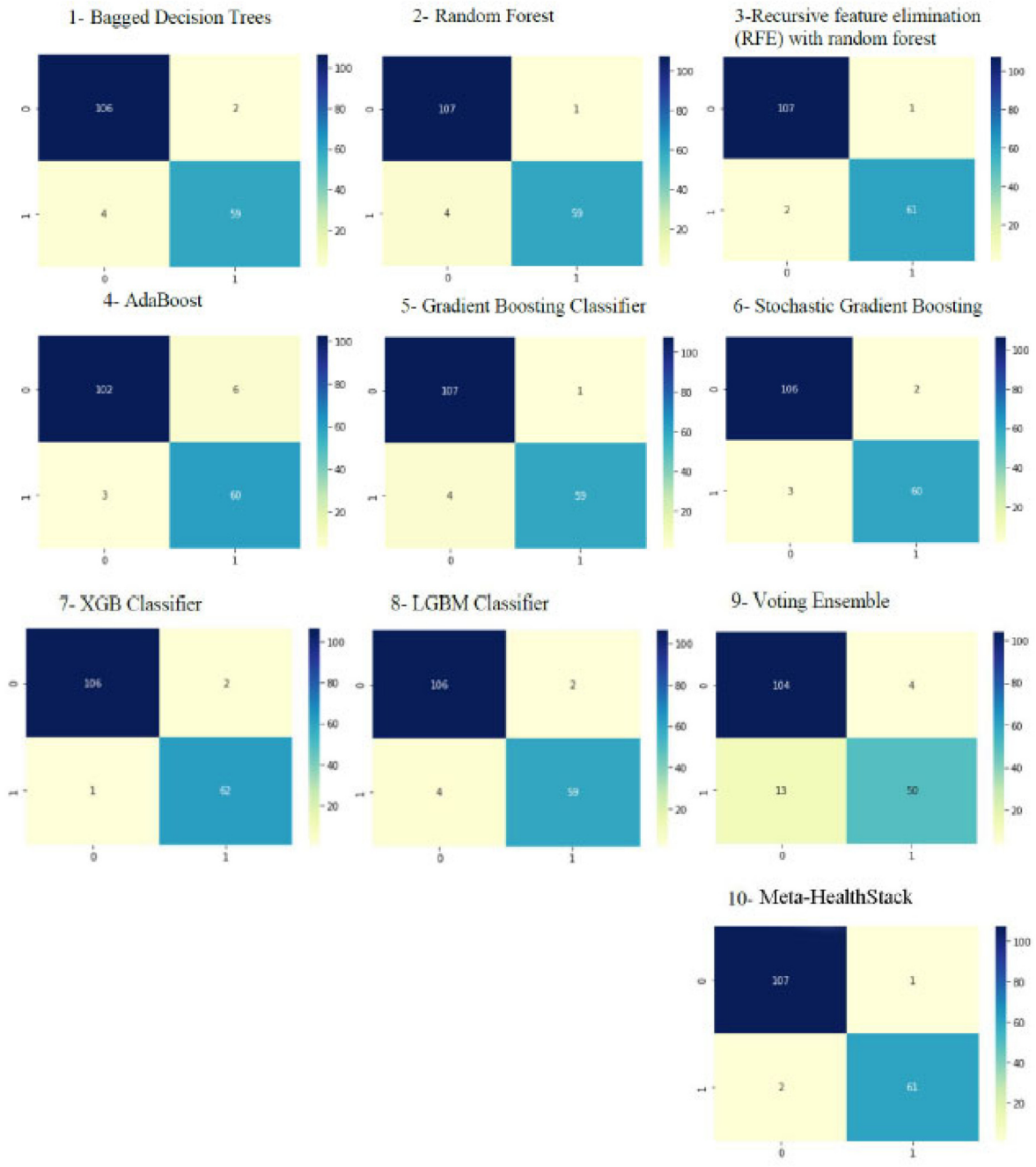

Fig. A.2. Each classification method confusion matrix.

Appendix B

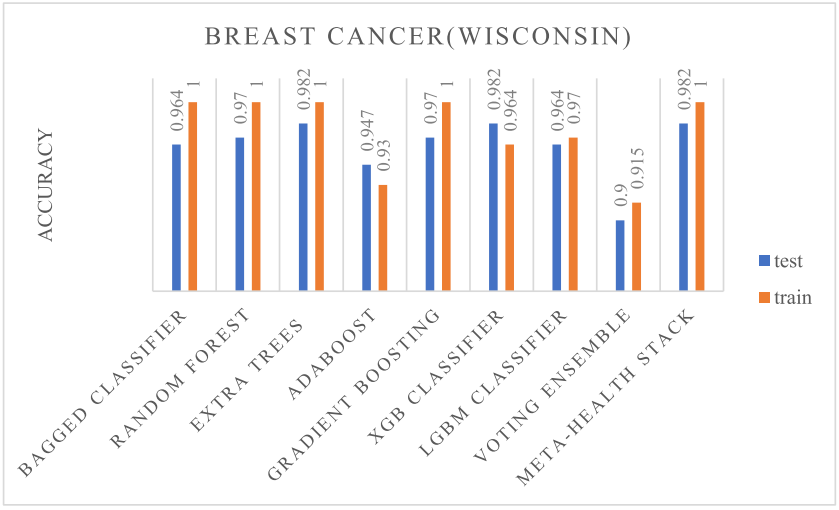

Fig. B.1. Methods output based on accuracy.

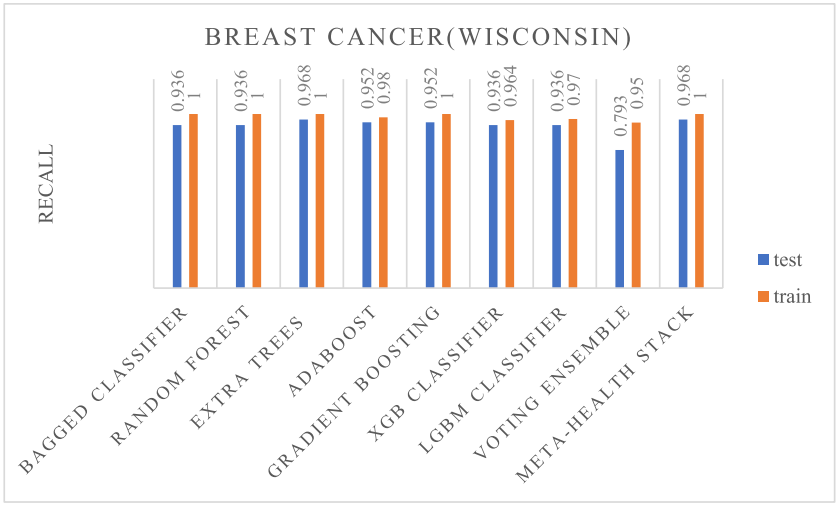

Fig. B.2. Methods output based on recall.

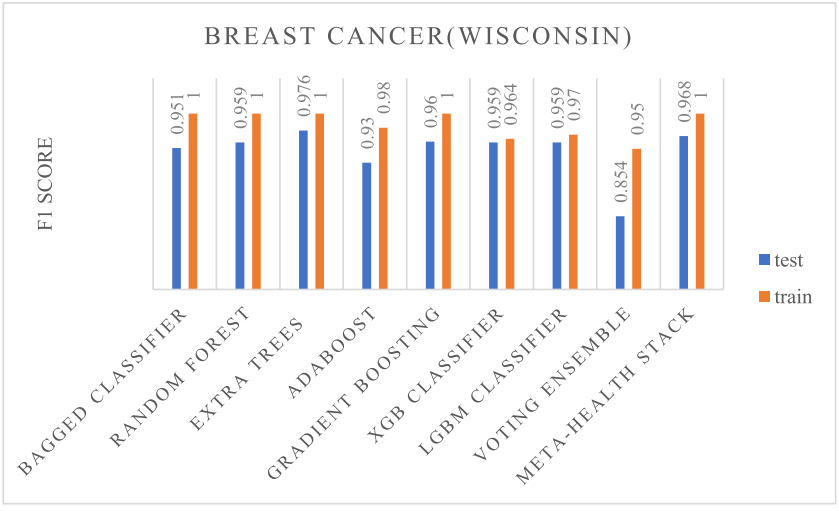

Fig. B.3. Methods output based on F1-score.

References

[1] R.L. Siegel, et al., Colorectal cancer statistics, 2020, CA: Cancer J. Clin. (2020).
[2] O. Ginsburg, et al., Breast cancer early detection: a phased approach to implementation, Cancer 126 (2020) 2379–2393.
[3] N. Emanet, et al., A comparative analysis of machine learning methods for classification type decision problems in healthcare, Decis. Anal. 1 (1) (2014) 1–20.
[4] A. Tartar, N. Kilic, A. Akan, Classification of pulmonary nodules by using hybrid features, Comput. Math. Methods Med. 2013 (2013).
[5] R.W. Brause, Medical analysis and diagnosis by neural networks, in: International Symposium on Medical Data Analysis, Springer, 2001.
[6] C. Arya, R. Tiwari, Expert system for breast cancer diagnosis: A survey, in: 2016 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2016.
[7] P. Israni, Breast cancer diagnosis (BCD) model using machine learning, Cancer Cells 1 (2019) 10.
[8] P. Singhal, S. Pareek, Artificial neural network for prediction of breast cancer, in: 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on, IEEE, 2018.
[9] A. Al Bataineh, A comparative analysis of nonlinear machine learning algorithms for breast cancer detection, Int. J. Mach. Learn. Comput. 9 (3) (2019) 248–254.
[10] B. Padmapriya, T. Velmurugan, Classification algorithm based analysis of breast cancer data, Int. J. Data Min. Tech. Appl. 6 (1) (2016) 43–49.
[11] M.K. Keleş, Breast cancer prediction and detection using data mining classification algorithms: a comparative study, Teh. Vjesn. 26 (1) (2019) 149–155.
[12] B. Panda, et al., A novel approach for breast cancer data classification using deep forest network, in: Intelligent and Cloud Computing, Springer, 2019, pp. 309–316.
[13] B. Zheng, S.W. Yoon, S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Syst. Appl. 41 (2013) 1476–1482.
[14] A. Kerhet, et al., A SVM-based approach to microwave breast cancer detection, Eng. Appl. Artif. Intell. 19 (7) (2006) 807–818.
[15] G. Sahebi, et al., GeFeS: A generalized wrapper feature selection approach for optimizing classification performance, Comput. Biol. Med. 125 (2020) 103974.
[16] M. Rostami, et al., Review of swarm intelligence-based feature selection methods, Eng. Appl. Artif. Intell. 100 (2021) 104210.
[17] E. Hancer, A new multi-objective differential evolution approach for simultaneous clustering and feature selection, Eng. Appl. Artif. Intell. 87 (2020) 103307.
[18] M.H. Memon, et al., Breast cancer detection in the iot health environment using modified recursive feature selection, Wirel. Commun. Mob. Comput. 2019 (2019).
[19] A.A. Said, et al., Classification based on clustering model for predicting main outcomes of breast cancer using hyper-parameters optimization, Int. J. Adv. Comput. Sci. Appl. 9 (12) (2018) 268–273.
[20] A. Kumara, R. Sushila, A.K. Tiwarib, Feature extraction and elimination using machine learning algorithm for breast cancer biological datasets, Int. J. Adv. Sci. Technol. 28 (20) (2019) 425–435.
[21] S.J. Pasha, E.S. Mohamed, Bio inspired ensemble feature selection (BEFS) model with machine learning and data mining algorithms for disease risk prediction, in: 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE, 2019.
[22] A. Ed-daoudy, K. Maalmi, Breast cancer classification with reduced feature set using association rules and support vector machine, NetMAHIB 9 (1) (2020) 34.
[23] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000.
[24] B. Krawczyk, G. Schaefer, M. Woźniak, A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification, Artif. Intell. Med. 65 (3) (2015) 219–227.
[25] M.-W. Huang, et al., SVM and SVM ensembles in breast cancer prediction, PLoS One 12 (1) (2017) e0161501.
[26] H.T.T. Thein, K.M.M. Tun, An approach for breast cancer diagnosis classification using neural network, Adv. Comput. 6 (1) (2015) 1.
[27] M.N. Haque, et al., Optimising weights for heterogeneous ensemble of classifiers with differential evolution, in: 2016 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2016.
[28] M. Abdar, V. Makarenkov, CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer, Measurement 146 (2019) 557–570.
[29] M.R. Basunia, et al., On predicting and analyzing breast cancer using data mining approach, in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020.
[30] P. Kumar, A. Gangal, S. Kumari, Prognosis of breast cancer by implementing machine learning algorithms using modified bootstrap aggregating, in: Innovations in Computational Intelligence and Computer Vision, Springer, 2020, pp. 561–569.
[31] P. Srimani, M.S. Koti, Medical diagnosis using ensemble classifiers-a novel machine-learning approach, J. Adv. Comput. 1 (2013) 9–27.
[32] M.M. Islam, et al., Breast cancer prediction: a comparative study using machine learning techniques, SN Comput. Sci. 1 (5) (2020) 1–14.
[33] M. Patrício, et al., Using resistin, glucose, age and BMI to predict the presence of breast cancer, BMC Cancer 18 (1) (2018) 29.
[34] S.A. Davidsen, M. Padmavathamma, Multi-modal evolutionary ensemble classification in medical diagnosis problems, in: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2015.
[35] D. Bardou, K. Zhang, S.M. Ahmad, Classification of breast cancer based on histology images using convolutional neural networks, IEEE Access 6 (2018) 24680–24693.
[36] B. Swiderski, et al., Novel methods of image description and ensemble of classifiers in application to mammogram analysis, Expert Syst. Appl. 81 (2017) 67–78.
[37] T. Tran, U. Le, Predicting breast cancer risk: A data mining approach, in: International Conference on the Development of Biomedical Engineering in Vietnam, Springer, 2017.
[38] H.M. Zolbanin, D. Delen, A.H. Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst. 74 (2015) 150–161.
[39] B. Pandey, et al., Evolutionary modular neural network approach for breast cancer diagnosis, Int. J. Comput. Sci. Issues 9 (1) (2012) 219–225.
[40] M.U. Salma, BAT-ELM: a bio inspired model for prediction of breast cancer data, in: 2015 International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT), IEEE, 2015.
[41] K. Munir, et al., Cancer diagnosis using deep learning: a bibliographic review, Cancers 11 (9) (2019) 1235.
[42] P. Mekha, N. Teeyasuksaet, Deep learning algorithms for predicting breast cancer based on tumor cells, in: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), IEEE.
[43] P. Gupta, S. Garg, Breast cancer prediction using varying parameters of machine learning models, Procedia Comput. Sci. 171 (2020) 593–601.
[44] UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 1992, [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29].
[45] R.E. Schapire, Explaining adaboost, in: Empirical Inference, Springer, 2013, pp. 37–52.
[46] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[47] G. Ke, et al., Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 3146–3154.
[48] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[49] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.
[50] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[51] N.L. Johnson, A.W. Kemp, S. Kotz, Univariate Discrete Distributions, Vol. 444, John Wiley & Sons, 2005.
[52] I. Guyon, et al., Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[53] R.A. Stine, Graphical interpretation of variance inflation factors, Amer. Statist. 49 (1) (1995) 53–56.
[54] J. Benesty, et al., Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, 2009, pp. 1–4.
[55] J.T. Kent, Information gain and a general measure of correlation, Biometrika 70 (1) (1983) 163–173.