Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 1561–1570

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

F-test feature selection in Stacking ensemble model for breast cancer prediction

Dhanya R, Irene Rose Paul, Sai Sindhu Akula, Madhumathi Sivakumar, Jyothisha J Nair

[a]*Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam - 690525,Kerala, India*

Abstract

Cancer data sets contains many details of patient information, out of which only a few attributes contribute in predicting the accurate stage of cancer. Certain attributes of the entire data set play a major role in deciding the type of cancer i.e. whether benign or malignant hence feature selection techniques are useful in such scenarios for retaining the relevant feature set. Moreover, in order to achieve our goal of predicting the accurate stage of cancer, we need an appropriate model which generally results in higher accuracy and ensemble model proves to be the best model for such scenarios. In this study, we are using the existing ensemble techniques along with a combination of supervised machine learning algorithms to develop a new model for breast cancer prediction. We are also using feature selection techniques to enhance the performance of the ensemble model. For this purpose, machine learning algorithms like Support Vector Machines, Naive Bayes, K-Nearest Neighbors, Logistics Regression and feature selection techniques like Variance threshold and f-test have been taken into consideration. To achieve higher accuracy for the ensemble model, bagging, boosting and stacking techniques are used.

Keywords: Ensemble model;feature selection;machine learning;bagging;boosting; stacking;

1. Introduction

In today 's world cancer has become one of the leading causes of death. According to the World Health Organization (WHO), results say that the deaths due to cancer in India are higher than in many countries, especially breast cancer. It is one of the leading causes of women mortality worldwide. Though cervical cancer was the most common cancer among Indian women, the incidence of breast cancer has surpassed cervical cancer as the leading cause of death in India. It is common among the younger age groups (in their thirties and forties) and mostly occurs in women aged from 30 to 69. There are many features to be taken into consideration while validating breast cancer which includes genetics, family history, lifestyle and so forth. The risk of occurrence of breast cancer may change with respect to change in any of these features. A regular diagnosis helps us to accumulate the variations and the impacts on a particular patient. Machine Learning is one of the fields which can be used for the prediction of breast cancer. It is the field of study that gives computers the capability to learn without being explicitly programmed. The process of learning begins with observations or data, such as direct experience or instruction, in order to look for patterns in data

10.1016/j.procs.2020.04.167

and make better decisions in the future based on the examples that we provide. In this study, we are focusing on the use of supervised machine learning algorithms that can apply what has been learned in the past to new data using labeled examples to predict future events. Ensemble method is a machine learning technique that combines several base models in order to produce one optimal model. Feature Selection is the process of selecting the features which contribute the most to the prediction variable or output thereby increasing efficiency of a model. In this study, we have used feature selection techniques along with ensemble techniques to build an optimal predictive model.

1.1. Literature Survey

There are already several models mentioned in the literature which help to predict breast cancer. The Gail model [1] uses non-genetic risk factors like age of the women during first child, number of present and past biopsies, race and so forth as features for the model. This model uses logistic regression for classification. Genetic factors can also attribute to the risk of cancer, which is not being recognized in this model. This is a drawback in this model. Based on the BRCA gene mutations, BRCAPRO model [2] predicts the patient's lifetime risk over the upcoming 5 years. Similar to Gail model, this model also considers ethnicity and both maternal and paternal family history. Bayes theorem comes into light for calculation of cancer probabilities. The Tyrer-Cuzick model [3] considers a wider range of features as input compared to the other models. Along with the above mentioned factors, this model also considers benign breast disease, BMI and genetic factors. This model uses Bayes theorem for the evaluation of risk. Though having more number of features in a model for prediction might seem desirable, it might sometimes over-fit the scenario and not bring us desirable results. Feature Selection algorithms are useful in such scenarios. The effect of feature selection algorithms has been studied in the scenario of cancer prediction as well as in other domains. In one of the works, a method call Kernel F-test Feature Selection algorithm [4] has been proposed for Breast cancer prediction. Kernel functions have been used in this method to increase the dimensions of the input features for clear separation of non-linearity. For all the features F-score values in the kernel space are calculated. The mean kernel F-score value is used as the threshold for selecting the features. SVM classifier is used for classification in this model. A study conducted in the domain of educational data mining has identified various combinations of machine learning algorithms and feature selection techniques [5] which perform well in the scenario of educational data mining. Some of the combinations include Decision trees & Correlation based feature selection, Voted perceptron & One Rule, Random Forest Classification & Gain Ratio. Another study [6] has proved that reducing the number of features on traditional machine learning classifiers enhances the performance as the computational complexity is reduced. It considers filter, wrapper and hybrid feature selection methods and it's affect on machine learning algorithms like Decision trees, Bayesian Networks and SVM. With a trial and error exhaustive search, using a comparative analysis the best features can be extracted out of a huge dataset. Forward and backward search, trial and error exhaustive search are a few methods used in the experiments with neural networks and combination of heuristics. According to the results of this study [7], SVM has a good accuracy of 97.38 and reduces the space from 32 to 6 which is better than other time consuming methods. A fusion of multi-classifiers and feature reduction transformation method on WBC, WDBC, WPBC has been applied. MLP and J48 with PCA is the best fusion for WBC [8]. Similarly for WDBC, SMO and MLP or Single Classifiers (SMO) and IBK. A study lightens the theory, removal of redundant features aids in the improvement of the performance of the classifier. A threshold is being set and the performance is compared on three different datasets based on a software metrics. In total there are eight metrics under which the software evaluates and it has been concluded that AUC parameter generally provides the better performance. The classifier still works better in most cases even when we remove 96 percent of the available software. Many scientists infer that machine learning is not just limited to one field [9]. A existing theory uses machine learning classifiers such as K-near neighbor, random forest using 10-fold cross validation, decision tree - c4.5 and Naive Bayes for early prediction of coronary heart events [10]. In another strategy [11] they have used MR images generated by CAD systems for Alzheimer's disease and have proposed a method based on multivariate approaches, handling the whole image as a single observation where segmentation of the gray matter region in the image is performed using Gaussian Mixture Models and extraction of the score vectors is done using Partial Least Squares. SVM classifier is used for the classification of the dataset and the proposed method is shown to have better performance than the existing models. In [12], lung cancer CT scan images are used to predict if the cancer is benign or malignant using AlexNet, a pre trained convolution neural network. They obtained an accuracy of 98 percent for the same. According to a study [13], with few features such as sugars, carbs, vitamin A and C, calories, the quality of a brand of juice is predicted. This proves

that choosing right features is important as it influences the result and hence care should be taken for feature selection when constructing models. Misclassified data and instances can influence the performance of any model and it is hard to figure them out when the data gets larger [14]. So this paper talks about how C - Support Vector classification filter works to identify misclassifed instances and thus remove them. Only the correctly classified instances would be later passed on to the learning algorithm which becomes our model. Performance is measured with accuracy and ROC as well as with other ensemble algorithms like Adaboost, Bagging and ensemble of SVM with adaboost and bagging filters. By applying this filter, the paper suggests that around 5 - 20 percent of misclassified data is being removed.

1.2. Proposed Method

In this paper we propose to develop an ensemble model which gives better predictive performance than the existing classifiers for breast cancer prediction. For this we have considered classifiers such as Naive Bayes, Support Vector Machine (SVM), Decision Tree, Multi-Layer Perceptron (MLP), Logistic Regression and K-nearest neighbours (KNN) and feature selection algorithms, namely f-test and variance thresholding.

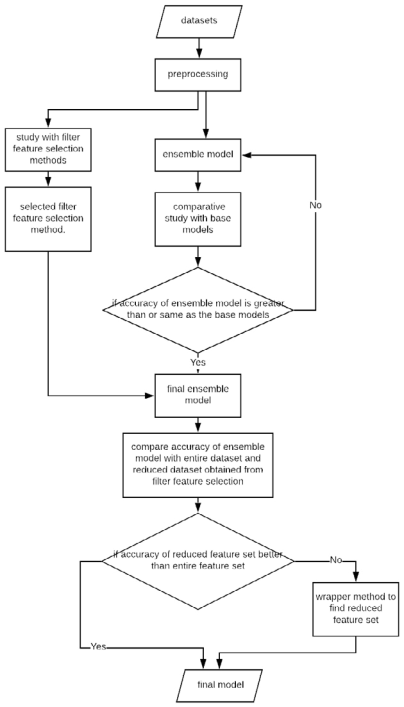

Fig. 1. Flowchart of the proposed system

1.3. Working of the Proposed Method

The datasets are first preprocessed. Then filter feature selection methods are applied on the preprocessed datasets and the method which gives better result is chosen and reduced feature set f is obtained using this filter feature selection method. An ensemble method is then developed with datasets as inputs. A comparative study is done to compare its performance with its base classifiers. If the performance is greater than or the same as its base classifiers, then the ensemble model is finalized. The accuracy of the final ensemble method with the entire feature set d is obtained. The accuracy of the final ensemble method with the reduced feature set f is obtained. Then both the accuracies are compared. If the ensemble model with reduced feature set give higher accuracy, then this model is finalized, else a

wrapper method is used to obtain reduced feature set and then ensemble model is applied on this feature set. Figure 1 illustrates the working of the proposed method.

1.4. Dataset

We have used openly available datasets like Breast Cancer Wisconsin (Diagnostic) DataSets from UCI Repository which consists of Breast cancer Wisconsin dataset, WDBC dataset and MicroRNA dataset from GIT repository. Breast cancer Wisconsin dataset consists of 699 instances with 11 attributes, namely Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class. The WDBC dataset contains features extracted from digitized image of a fine needle aspirate of a breast mass which describe the characteristics of the cell nuclei in the image. This dataset consists of 569 instances with 32 attributes The MicroRNA dataset contains information on tumour microRNA expression in 133 samples of primary breast cancer tissue which considers 1928 tumor miRNA signatures for the discrimination of breast cancer and the intrinsic molecular subtypes.

1.5. Pre-processing

We preprocessed the datasets obtained using two methods:

1. Removal of missing values - Here we replaced all the missing values with the mean values of the columns as removal of rows with missing values for small datasets can lead to shortage of training data.
2. Normalization - In dataset 1 (wisconsin) and dataset 2 (wdbc), the range of values belonging to the columns differed from one another. Therefore normalization was done to change the values of numeric columns in the dataset to be within a common scale, without distorting differences in the ranges of values or losing information. All the values were brought within the range 0-1 after normalization. The formula for normalization is as follows:

$$x' = (x - x_{min})/(x_{max} - x_{min}) \tag{1}$$

where
x_{min} is the minimum value of a feature
x_{max} is the maximum value of a feature
x' is the normalized value

1.6. Filter Feature Selection

1. We have used two feature selection techniques f-test and variance thresholding which are filter methods. The features are selected on the basis of their scores in various statistical tests for their relation with the outcome variable and not with regards to the machine learning algorithms.

1.6.1. F-Test

F-test is a statistical test that gives an f-score by calculating the ratio of variances. In this paper we have used f-test in one-way Analysis of Variance (ANOVA) which calculates the ratio of variance between groups and the variance within a group for a feature. The groups in this case are the instances with the same target value. Greater value of f-score means that the distances within the groups are less and distances between the groups are more. In this feature selection method of ANOVA using f-test, the features are ranked based on higher values of f-score. The f-score in this method is given by:

$$f - score = variancebetweengroups/variancewithingroups \tag{2}$$

where
variance between groups is the variance between groups indicated by the target feature
variance within group is the sum of variances within each group
x' is the normalized value

1.6.2. Variance Thresholding

It is a filter feature selection technique motivated by the idea that low variance features contain less information. It calculates the variance of each feature and removes those with variance less than a given threshold. This feature selection algorithm looks only at the features (X), not the desired outputs (y), and can thus be used for unsupervised learning. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.

2. In order to assess the performance of machine learning algorithms on the similar subsets of features obtained from filter methods for the given datasets, we have considered the following algorithms - Logistic Regression, Naive Bayes, Decision Tree, Support Vector Machine, K-nearest neighbours and Multi-layer Perceptron.
3. We used f-test feature selection to select the k highest ranking features where k is the number of features required. k is a hyper-parameter. We have given values starting from 1 to total number of features in the dataset for k.
4. In case of variance thresholding the hyper-parameter is the threshold value of the variance. We have given values ranging from the smallest variance value to the highest variance value obtained for each of the features in a particular dataset.
5. After the feature subsets are obtained, we have used them to train each of the machine learning algorithms mentioned above to obtain the accuracy corresponding to each subset.
6. A graph showing accuracy v/s k value has been plotted for f-test and accuracy v/s threshold value has been plotted for variance thresholding.
7. The k-value and the threshold value for f-test and variance thresholding respectively giving the highest accuracy obtained after applying each of the machine learning algorithms is considered.

1.7. Ensemble Methods

1. We have used three ensemble techniques in this project namely bagging, boosting and stacking.

1.7.1. Bagging

Bagging is the application of the Bootstrap procedure on a high-variance machine learning algorithm, typically decision trees i.e. it uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression. Bagging helps in reducing high variance and thereby preventing over-fitting.

1.7.2. Boosting

It is a technique that focuses on fitting sequentially multiple weak learners models that are only slightly better than random guessing, such as small decision trees, in a very adaptive way: each model in the sequence is fitted giving more importance to those observations in the dataset that were badly handled by the previous models in the sequence. It mainly focuses on reducing bias. The two most widely used boosting algorithms are adaboost and gradient boosting. In this project we have considered adaboost algorithm for performing boosting.

1.7.3. Stacking

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained on the complete training set, then the meta-model is trained with the outputs of the base level models as features.

2. In this project we have considered those base classifiers which give less accuracy for the entire dataset. We have used these to build the ensemble models until we could come up with an ensemble model which could give a better or same accuracy consistently for all the three datasets.
3. For bagging and boosting we have considered three base models - SVM, SVM with RBF kernel and naive bayes. We have built bagging and boosting models on each of the above mentioned algorithms individually.
4. For stacking we have considered SVM, KNN and Naive Bayes as the base classifiers for the first level of prediction. The predictions given by these algorithms are the input features for the second level. For the second level of prediction, we have considered logistic regression as the base classifier.

1.8. Comparative Study

- The objective of this module is to do a comparative study of the ensemble model with the base classifiers used to build them.
- If the ensemble model proposed does not give the same or better accuracy then the work flow of the project goes back to ensemble model module where further ensemble techniques are experimented.
- Correspondingly we perform filter feature selection on the dataset and the dataset size vs accuracy graph is plotted.
- We then compare the accuracies given by the entire feature set and the reduced feature set with the ensemble model.

1.9. Wrapper Feature Selection

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the model, we decide to add or remove features from a subset. These methods are usually computationally very expensive. The objective of this module is to do feature selection using wrapper methods if the ensemble model does not give a better accuracy for the reduced feature set obtained after applying filter feature selection methods on the datasets.

2. Experimental Results

For variance threshold, we have applied normalization as a preprocessing step in order to bring the features to a common scale whereas, f-test does not require normalization. Tables 1, 2 and 3 compare the accuracies of various base classifiers before normalization and after normalization for Wisconsin, WDBC and microRNA datasets respectively.

2.1. Accuracy before and after normalization on datasets

classifier	before normalization	after normalization
SVM	97.14	87.86
Logistic Regression	96.43	85.71
KNN	97.86	87.86
Naive Bayes	95.71	86.43
Decision Tree	93.57	90.71
MLP	60.71	60.71

Table 1. Classifier accuracy before normalization and after normalization on Wisconsin Dataset

classifier	before normaliztion	after normalization
SVM	95.61	58.77
Logistic Regression	95.61	71.93
Naive Bayes	92.98	82.46
KNN	93.86	93.86
Decision Tree	91.23	93.86
MLP	92.11	58.77

Table 2. Classifier accuracy before and after normalization on WDBC Dataset

classifier	before normalization	after normalization
SVM	100.00	88.89
Logistic Regression	100.00	100.00
KNN	100.00	100.00
Naive Bayes	100.00	100.00
Decision Tree	95.00	100.00
MLP	87.50	100.00

Table 3. Classifier accuracy before normalization and after normalization on MicroRNA Dataset

classifier	accuracy	maximum accuracy
SVM	87.86	88.57
Logistic Regression	85.71	90.00
KNN	87.86	93.57
Naive Bayes	86.43	86.43
Decision Tree	90.71	92.14
MLP	60.71	88.57

Table 4. Comparison Of Base accuracy with highest accuracy obtained after variance thresholding for Wisconsin Dataset

Tables 4, 6 and 8 show the maximum accuracies for base classifiers obtained after applying variance threshold feature selection for Wisconsin, WDBC and MicroRNA datasets respectively. Tables 5, 7 and 9 show the maximum accuracies for base classifiers obtained after applying f-test feature selection for Wisconsin, WDBC and MicroRNA datasets respectively. From these tables it is seen that applying feature selection can help us obtain the same accuracies as the entire feature set or better accuracies when reduced subset of features is used. It can also be seen that at some points the accuracies can go below the accuracies of the entire feature set. This implies that if the relevant features

classifier	accuracy	maximum accuracy
SVM	97.14	97.14
Logistic Regression	96.43	97.14
KNN	97.86	97.86
Naive Bayes	95.71	97.14
Decision Tree	93.57	95.71
MLP	60.71	96.43

Table 5. Comparison Of Base accuracy with highest accuracy obtained after f-test for Wisconsin Dataset

classifier	accuracy	maximum accuracy
SVM	58.77	70.18
Logistic Regression	71.93	71.93
KNN	93.86	91.23
Naive Bayes	82.46	94.74
Decision Tree	93.86	92.98
MLP	58.77	93.86

Table 6. Comparison Of Base accuracy with highest accuracy obtained after variance thresholding for WDBC Dataset

classifier	accuracy	maximum accuracy
SVM	95.61	96.49
Logistic Regression	95.61	95.61
Naive Bayes	92.98	96.49
KNN	93.86	93.86
Decision Tree	91.23	95.61
MLP	92.11	95.61

Table 7. Comparison Of Base accuracy with highest accuracy obtained after f-test for WDBC Dataset

classifier	accuracy	maximum accuracy
SVM	88.89	100.00
Logistic Regression	100.00	100.00
KNN	100.00	100.00
Naive Bayes	100.00	100.00
Decision Tree	100.00	100.00
MLP	100.00	100.00

Table 8. Comparison Of Base accuracy with highest accuracy obtained after variance thresholding for MicroRNA Dataset

classifier	accuracy	maximum accuracy
SVM	100.00	100.00
Logistic Regression	100.00	100.00
KNN	100.00	100.00
Naive Bayes	100.00	100.00
Decision Tree	95.00	100.00
MLP	87.50	100.00

Table 9. Comparison Of Base accuracy with highest accuracy obtained after f-test for MicroRNA Dataset

classifier	accuracy	threshold accuracy
SVM	87.86	82.14
Logistic Regression	85.71	75.71
KNN	87.86	85.00
Naive Bayes	86.43	80.00
Decision Tree	90.71	81.43
MLP	60.71	61.43

Table 10. Comparison Of Base accuracy with accuracy obtained by taking mean of variances as threshold for Wisconsin Dataset

classifier	accuracy	maximum accuracy
SVM	97.14	97.14
Logistic Regression	96.43	96.43
KNN	97.86	94.29
Naive Bayes	95.71	95.00
Decision Tree	93.57	93.57
MLP	60.71	60.71

Table 11. Comparison Of Base accuracy with accuracy obtained by taking mean of f-scores as threshold for Wisconsin Dataset

classifier	accuracy	threshold accuracy
SVM	58.77	59.65
Logistic Regression	71.93	71.93
KNN	93.86	93.86
Naive Bayes	82.46	94.74
Decision Tree	93.86	91.23
MLP	58.77	58.77

Table 12. Comparison Of Base accuracy with accuracy obtained by taking mean of variances as threshold for WDBC Dataset

classifier	accuracy	maximum accuracy
SVM	95.62	93.86
Logistic Regression	95.61	93.86
Naive Bayes	92.98	92.98
KNN	93.86	93.86
Decision Tree	91.23	93.86
MLP	92.11	58.77

Table 13. Comparison Of Base accuracy with accuracy obtained by taking mean of f-scores as threshold for WDBC Dataset

classifier	accuracy	threshold accuracy
SVM	88.89	100.00
Logistic Regression	100.00	100.00
KNN	100.00	100.00
Naive Bayes	100.00	100.00
Decision Tree	100.00	100.00
MLP	100.00	100.00

Table 14. Comparison Of Base accuracy with accuracy obtained by taking mean of variances as threshold for MicroRNA Dataset

are removed it leads to a decrease in accuracy whereas if relevant features are chosen it can lead to an increase in the performance. Dataset size v/s accuracy graphs were plotted as a study of the effect of feature selection on various base classifiers. It is seen that f-test improves the accuracies of the base classifiers in many cases whereas variance thresholding either retains or reduces the accuracies of the base classifiers for each of the datasets.

While doing feature selection we need to give the threshold value as input parameter in the case of variance thresholding whereas the number of features required should be given for f-test. For this purpose we came up with a method to calculate the average of all the variances obtained for each feature in variance thresholding and the average of all the f-score values obtained for each feature in f-test. We then used this mean value as a threshold value for both the feature selection technique. Tables 10, 12 and 14 show the accuracies for base classifiers without applying feature selection and those obtained after applying variance threshold feature selection with the mean value set as the threshold

classifier	accuracy	maximum accuracy
SVM	100.00	100.00
Logistic Regression	100.00	100.00
KNN	100.00	100.00
Naive Bayes	100.00	100.00
Decision Tree	95.00	95.00
MLP	87.50	87.50

Table 15. Comparison Of Base accuracy with accuracy obtained by taking mean of f-scores as threshold for MicroRNA Dataset

classifier	base accuracy	boosting	bagging
Naive Bayes	92.98	94.74	92.11
SVM linear	95.61	95.61	95.61
SVM rbf	58.77	58.77	58.77

Table 17. Comparison Of accuracies obtained after applying Bagging and Boosting for WDBC Dataset

Dataset	Accuracy without	Acuracy with
Wisconsin	97.14	97.14
WDBC	96.49	97.37
MicroRNA	100.00	100.00

Table 19. Accuracy for Stacking on all the three datasets with and without feature selection

classifier	base accuracy	boosting	bagging
Naive Bayes	95.71	95.00	95.71
SVM linear	97.14	97.86	97.14
SVM rbf	95.0	60.71	94.29

Table 16. Comparison Of accuracies obtained after applying Bagging and Boosting for Wisconsin Dataset

classifier	base accuracy	boosting	bagging
Naive Bayes	100.00	100.00	90.00
SVM linear	100.00	100.00	100.00
SVM rbf	88.88	87.50	87.50

Table 18. Comparison Of accuracies obtained after applying Bagging and Boosting for MicroRNA Dataset

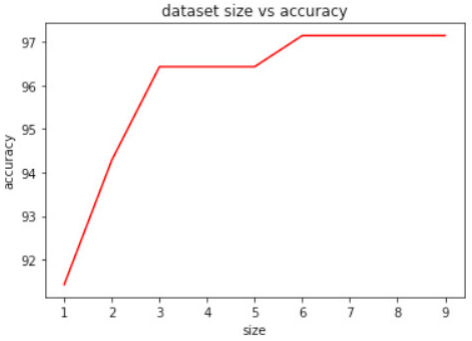

Table 20. Accuracy vs dataset size (stacking with f-test feature selection) for Wisconsin dataset

for Wisconsin, WDBC and MicroRNA datasets respectively. Similarly Tables 11, 13 and 15 show the accuracies of base classifiers and those obtained after applying f-test feature selection with the mean value set as the threshold for Wisconsin, WDBC and MicroRNA datasets respectively. From these tables it can be seen that setting the mean value as threshold does not give a good accuracy consistently. In some cases it decreases the accuracy considerably whereas in others it mostly gives the same accuracy as without using feature selection. Hence this method is not an efficient way to tackle the problem of threshold value calculation. As mentioned above, since f-test either increases or retains the same accuracy as base classifiers, this feature selection method has been used for further experiments

As mentioned in the proposed system we have used three ensemble techniques for building up a good model for breast cancer prediction. Based on the literature study we have selected three classifiers - naive bayes, SVM and SVM with RBF kernel. Since Bagging and Boosting are homogeneous ensemble techniques we have obtained the results as shown in Tables 16, 17 and 18 after applying them on each of the classifiers mentioned above. In Bagging and Boosting it can be seen that the ensemble model retains the accuracy of the base classifier in most of the cases whereas it decreases in some cases. Hence there is no significant improvement in the performance of these models.

Next we applied stacking- which is a heterogenous ensemble technique - using classifiers such as KNN, SVM with linear kernel, Naive Bayes and Logistic Regression and the result obtained is shown in Table 19 which is as follows: The stacking model obtained a better accuracy for WDBC dataset. It gave an accuracy equal to the highest accuracy obtained with each of the base classifiers in the case of Wisconsin and MicroRNA dataset. Later we applied f-test feature selection and examined the accuracy of the subsets obtained using stacking to check if the accuracy can be improved with feature selection. For WDBC dataset a better accuracy than that given without using feature selection was obtained. For the other two datasets the accuracy continued to remain the same even after applying f-test feature selection on them. Figures 1, 2 and 3 show the graphs of accuracy vs the size of the dateset when stacking with f-test feature selection is used for Wisconsin, WDBC and MicroRNA datasets respectively.

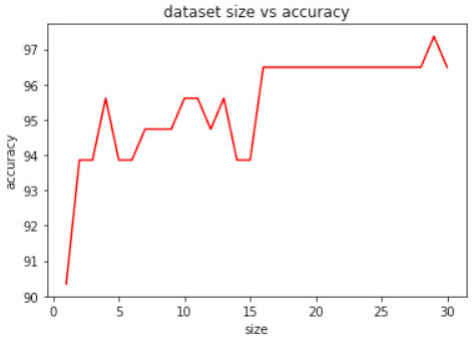

Fig. 2. Accuracy vs dataset size (stacking with f-test featureselection) for WDBC dataset

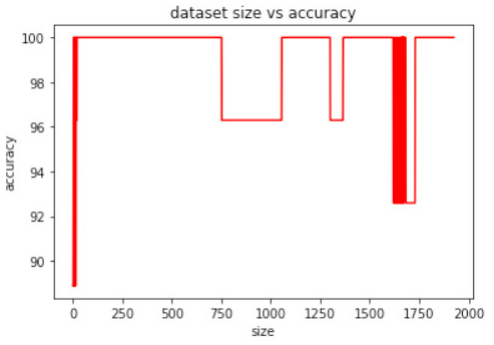

Fig. 3. Accuracy vs dataset size (stacking with f-test feature selection) for MicroRNA dataset

Based on the various experiments mentioned above, it is seen that stacking ensemble model with f-test feature selection gives the best accuracy. The feature set obtained after applying f-test on the three datasets appears to give a better accuracy for the six base classifiers used i.e. SVM, MLP, Logistic Regression, Decision Tree, Naive Bayes and KNN to assess the performance of the feature selection methods. F-test retains or improves the accuracies of the classifiers for some subsets of the datasets. This is because F-test calculates the ratio of variance between two groups and the within the group for a feature. Here the groups refer to the target classes. Hence it provides high scores for those features which better distinguish the classes. Also though feature selection is effective in general, it will not improve the performance of the model if the feature selection algorithm excludes features relevant for classification.

Of all the ensemble techniques used for breast cancer prediction stacking ensemble model which uses a combination of Naive Bayes, SVM (linear kernel), KNN and Logistic Regression gives a consistently better accuracy for WDBC dataset and retains the same accuracy as that of the base classifiers for Wisconsin and MicroRNA datasets whereas the other methods seems to perform less efficient than stacking. The is partly due to the fact that stacking is a heterogeneous ensemble technique and combines the results from various machine learning models rather than a single type of model to perform prediction. Also, in stacking the output of all the base models is fed into a meta-classifier which then makes the prediction based on those inputs. This shows that learning takes place even at the meta-level for stacking unlike the majority voting criteria used in bagging and boosting.

3. Conclusion

In this paper we can see that we obtain the same accuracy or even better accuracy for a subset of the entire feature set. From the various graphs we can conclude that not all the features are necessary for breast cancer prediction and feature selection helps in building an efficient model in such scenarios. Therefore we have used two filter feature selection techniques F-test and variance thresholding to obtain a reduced feature set for better prediction out of which f-test has given better results.

We have considered three ensemble techniques - bagging, boosting and stacking out of which stacking has turned out be a better predictive model for breast cancer prediction.

In the end, we proposed to perform wrapper feature selection methods along with the final ensemble model which would help in improving the ensemble accuracy if the filter method fails to produce a subset of features. But the use of f-test feature selection which is a filter method, led to an improvement in the accuracy of Stacking for WDBC dataset and for wisconsin and MicroRNA dataset as it gives the same accuracy for a reduced feature set thereby building an efficient and faster model. Also wrapper methods being computationally intensive will take a considerable amount of time in producing the reduced feature set for an ensemble model like Stacking. Hence we can conclude that stacking ensemble method with f-test feature selection is an effective and reliable way to predict breast cancer for the above mentioned datasets.

References

[1] Gail MH1, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, An-derson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, NormanSA, Simon MS, Spirtas R, Ursin G, Bernstein L., Projecting individu-alized absolute invasive breast cancer risk in African American women,J Natl Cancer Inst. 2007 Dec 5;99(23):1782-92. Epub 2007 Nov 27.

[2] Berry DA, Iversen ES Jr, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT, Hilsenbeck SG, Rubinstein WS, Hughes KS, Parmigiani G. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. J Clin Oncol. 2002 Jun 1;20(11):2701-12.

[3] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004 Apr 15;23(7):1111-30. Erratum in: Stat Med. 2005 Jan 15;24(1):156.

[4] Jaganathan P., Rajkumar N., Nagalakshmi R. (2011) A Kernel Based Feature Selection Method Used in the Diagnosis of Wisconsin Breast Cancer Dataset. In: Abraham A., Lloret Mauri J., Buford J.F., Suzuki J., Thampi S.M. (eds) Advances in Computing and Communications. ACC 2011. Communications in Computer and Information Science, vol 190. Springer, Berlin, Heidelberg.

[5] Zaffar, Maryam Ahmed, Manzoor Savita, K S Sajjad, Syed. (2018). A Study of Feature Selection Algorithms for Predicting Students Academic Performance. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090569.

[6] Kourou, Konstantina, P. Exarchos, Themis, Exarchos, Konstantinos, Karamouzis, Michalis, Fotiadis, Dimitrios. (2014). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 13.10.1016 /j.csbj.2014.11.005.

[7] Zheng, Bichen, Yoon, Sang Won, S. Lam, Sarah. (2013). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications. 41. 1476–1482. 10.1016/j.eswa.2013.08.044.

[8] I. Salama, Gouda, Abdelhalim, mohamed b, Zeid, Magdy. (2012). Experimental comparison of classifiers for breast cancer diagnosis. 180-185. 10.1109/ICCES.2012.6408508.

[9] H. Wang, T. M. Khoshgoftaar, and J. Van Hulse,A Comparative Study of Threshold-Based Feature Selection Techniques, 2010 IEEE International Conference on Granular Computing, pp. 499·504, 2010.

[10] R. Ani, Augustine, A., Akhil, N. C., and Dr. Deepa Gopakumar O. S., "Random Forest Ensemble Classifier to Predict the Coronary Heart Disease Using Risk Factors", in Proceedings of the International Conference on Soft Computing Systems, 2016.

[11] J. J. Nair and N. Mohan, "Alzheimer's disease diagnosis in MR images using statistical methods," 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2017, pp. 1232-1235.

[12] Sathyan H, Panicker J.V,"Lung Nodule Classification Using Deep ConvNets on CT Images",9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018

[13] N. Cheryl, A. Jaya, S. Krishnapriya and J. J. Nair, "Commercial Mango Juice Classification Using Fuzzy Logic," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 888-891.

[14] Thongkam J., Xu G., Zhang Y., Huang F. (2008) Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction. In: Ishikawa Y. et al. (eds) Advanced Web and Network Technologies, and Applications. APWeb 2008. Lecture Notes in Computer Science, vol 4977. Springer, Berlin, Heidelberg