ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 9 (2022) No.3, pp.129-145
DOI:10.33168/LISS.2022.0310

An Evaluation Study on the Predictive Models of Breast Cancer Risk Factor Classification

Wen San Yee[1], Hu Ng[1+], Timothy Tzen Vun Yap[1], Vik Tor Goh[2], Keng Hong Ng[1], Dong Theng Cher[3]

[1] Faculty of Computing and Informatics, Multimedia University, Malaysia
[2] Faculty of Engineering, Multimedia University, Malaysia
[3] SIRIM Berhad, Malaysia

Abstract. This research is intended to explore and evaluate various predictive models for the classification performance of breast cancer risk factors. First, data acquisition is being carried out to obtained three datasets from Breast Cancer Surveillance Consortium (BCSC). After that, data integration is performed to combine the datasets into one. Then, data preprocessing is performed to do data cleaning. Feature selection is executed to eliminate unrelated attributes. Data resampling is applied to resolve imbalanced data. Four classifiers namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) are used in classifying the risk factors of breast cancer. These four classifiers undergo training and testing data with 80-20, 70-30, and 60-40 train test splits. RF performs the best performance with 82% of accuracy at 80-20 train test split.

Keywords: breast cancer, boruta feature selection, data resampling, logistic regression, random forest, support vector machine, multilayer perceptron.

1. Introduction

Breast Cancer, the utmost commonly diagnosed cancer affecting women throughout the world. World Health Organization (World Health Organization, 2021) stated that there were approximately 2.3 million breast cancer cases and 685,000 deaths worldwide in 2020. This disease can also affect men. Breast cancer occurs when the cells from breast lobules turn abnormal by replicating uncontrollably. These cancerous cells from breast tissue will start to invade and spread through the surroundings of the body. Hence, early detection and diagnosis of breast cancer have higher chances of successful treatment and thus improving survival rates.

The involvement of Machine Learning (ML) has a huge contribution in the medical field because of its outstanding performance in estimating consequences. Moreover, the diagnosis of breast cancer will be more efficient and have a higher accuracy of outcomes by implementing ML methods. Therefore, developing a reliable and accurate ML model for making decisions regarding the risk factors of breast cancer is essential.

Machine learning models aim to identify the effective attributes and determine the relationship between them. These models are often used for prediction, estimation, and additionally for determining a method to design a good model which will be learned through experience to enhance its performance. In this research, our focus is on classifying the breast cancer risk factors with the approach of machine learning. This research addressed the scope by evaluating the performance of machine learning algorithms using publicly available data on the Breast Cancer Surveillance Consortium (BCSC) Breast Cancer Surveillance Consortium, 2021). By training the data model about the risk factors of breast cancer, the ML model can predict if the individual has a prior breast cancer diagnosis. Hence, early prevention actions and treatments can be made.

The objectives of this paper are to explore various predictive models to classify risk factors of breast cancer and to evaluate the performance of machine learning models.

2. Related Works

According to the study by Williams et al. (2015), Naive Bayes and J48 decision tree algorithms were the two data mining techniques used to predict breast cancer risks. The authors used the LASUTH dataset which is collected from the Cancer Registry of Lagos State University Teaching Hospital (LASUTH) in Nigeria to classify breast cancer. They had divided the breast cancer risk factors into two groups: changeable factors and non-changeable factors. Their experimental research results showed that the J48 decision tree is a better model in predicting breast cancer risks as compared to Naive Bayes.

Multiple studies have been conducted by using Wisconsin Breast Cancer datasets (Wolberg and Mangasarian, 1990) to predict breast cancer risks. Asri et al. (2016) demonstrated that Support Vector Machine (SVM) obtained the highest accuracy of 97.13%, which outperformed Decision Tree (C4.5), Naive Bayes (NB), and K-Nearest Neighbors (KNN).

Shajahaan et al. (2013) studied the behaviours of Random tree, Iterative Dichotomizer (ID3), Classification and Regression Tree (CART), C4.5, and Naive Bayes (NB) to predict the occurrence of cancer. Their research study revealed that the Random tree achieved 100% accuracy.

Lavanya & Rani (2013) had proposed a new hybrid method whereby CART decision tree classifier was combined with clustering and feature selection (FS) to enhance the accuracy of the classifier. The authors revealed that cascading classification with data mining algorithms can enhance classification accuracy.

Chaurasia et al. (2018) presented three popular data mining techniques: NB, Radial basis function (RBF) Network, and J48 algorithm in breast cancer risk factors detection. Their experimental results demonstrated that NB has the best performance with the highest accuracy of 97.36%.

Iqbal et al. (2019) employed four classifiers which are Multilayer Perceptron (MLP), SVM, KNN, and Random Forest (RF) to identify the most effective predictors for breast cancer prediction. Their results prove that RF is the best classifier for their research.

In a study conducted by Kabir & Ludwig (2018), the researchers had implemented classifiers of Decision Tree (DT), RF, and Extreme Gradient Boosting (XGBoost) to model the BSSC breast cancer dataset. As the dataset was imbalanced, the authors applied different resampling techniques on the training data. Their experimental results demonstrated notable enhancement when resampling methods was applied. Table 1 shows the summary of the highest accuracy obtained by various research groups.

3. Research Methodology

The flow of this research is shown in Figure 1. There are four major stages required to be performed before classification which are data acquisition, data processing, data cleaning, and feature selection.

Table 1: Summary of the highest accuracy obtained by various research groups.

Research group	Database	Classifier model	Accuracy (%)
Williams et al. (2015)	LASUTH breast cancer	J48 Decision Trees	94.2
Asri et al. (2016)	Wisconsin Breast Cancer	SVM	97.13
Shajahaan et al. (2013)		Random Tree	100
Lavanya & Rani (2013)		CART decision tree	98.71
Chaurasia et al. (2018)		Naïve Bayes	97.36
Naveed et al. (2019)		Random Forest	99.26
Kabir & Ludwig (2018)	BCSC	XGBoost with ENN	91.49

3.1. Data acquisition

The risk factor dataset chosen for this research was taken from the Breast Cancer Surveillance Consortium (BCSC) (Ballard-barbash et al. (2017). This dataset was developed from between January 2005 and December 2017. It contains 1,522,340 records of information. The database was gathered from seven mammography registries participating in the National Cancer Institute-funded Breast Cancer Surveillance Consortium (Ballard-barbash et al. (2017). Each registry collected data of patient's tumour characteristics, breast cancer diagnoses, demographic and clinical reports, and also mammogram evaluation data that represents the population of women in the United States undergoing mammography Sickles (2005). Lehman (2017) stated that women are required to complete a questionnaire which contains questions related with personal breast cancer history, menopausal status, and self-reported symptoms at each visit for every registry. Table 2 shows the 13 attributes available from the dataset.

3.2. Data integration

As the dataset was split into three zipped files which contain all the same basic information, but only the year of the record is different. Each zipped files contains one comma-separated values (CSV) file. All files are merged into one main CSV file.

3.3. Data cleaning

The research work checked if the dataset contains any missing values or duplicated values. Missing data or duplicated data will be removed. Additionally, the dataset contains undefined variables which are represented by value 9. Data containing unknown value 9 was excluded from the dataset to ensure its reliability.

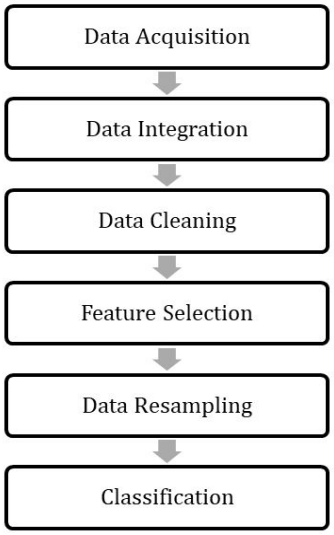

Fig. 1: Project stages framework.

Table 2: Exploration of dataset attributes.

Variable Name	Description
year	Calendar year of observation
age_group_5_years	Age (years) in 5 year groups
race_eth	Race/ethnicity
first_degree_hx	History of breast cancer in a first degree relative
age_menarche	Age (years) at menarche
age_first_birth	Age (years) at first birth
BIRADS_breast_density	BI-RADS breast density
current_hrt	Use of hormone replacement therapy
menopaus	Menopausal status
bmi_group	Body mass index (kg/m2)
biophx	Previous breast biopsy or aspiration
breast_cancer_history	Prior breast cancer diagnosis
count	Frequency count of this combination of covariates

3.4. Feature selection

The feature selection method used is known as Boruta which makes use of the BorutaPy package. Feature selection plays an important role in predicting a class to maximize performance by removing irrelevant features that cause unnecessary noise in the data which will then result in poor model accuracy. Boruta is a type of wrapper algorithm that is built around a classifier known as Random Forest (RF) machine learning model.

Boruta captures all the important and interesting features that affect the dependent variable (breast cancer detection), with respect to being independent of other variables. By setting up a threshold, it can identify which features are not important with the prediction. Values equal to 0.00 will be discarded as it seems to be an irrelevant feature by Boruta. Figure 2 shows the Boruta score of breast cancer risk factors.

As observed in Figure 2, age group and previous breast biopsy or aspiration have a perfect score of 1.00. This indicates that both features are very related to prior breast cancer.

	Features	Score
1	age_group	1.00
10	biophx	1.00
0	year	0.89
2	race_eth	0.78
5	age_first_birth	0.67
6	BIRADS_breast_density	0.56
9	bmi_group	0.44
8	menopause	0.33
4	age_menarche	0.22
7	current_hrt	0.11
3	first_degree_hx	0.00

Fig. 2: Boruta score for risk factor dataset.

3.5. Data resampling

For a set of imbalanced data, one common problem created as most of the data would fall into a certain class, which is the majority class, while the minority class has insufficient data points to make a valid comparison. This situation could lead to poor performance for data model training. The data resampling method chosen in this

research is Synthetic Minority Oversampling Technique (SMOTE). Figure 3 shows the count plot on the target variable of imbalanced data, whereby the value of 0 was having 292,132 and the value of 1 was 52, 932.

SMOTE is an oversampling approach that selects the nearest neighbors in a given feature space, generates a line to separate the examples, then produces new samples on the line. At first, synthetic samples from the minority class were generated. To balance the class distribution, the minority class examples were randomly increased by replication. SMOTE will loop through the synthesis samples and at each iteration, a positive class instance was selected at random. For this research, SMOTE technique with a random state of 10 and K-nearest neighbors of 5 was applied to resample the outcome. In the end, the value of 1 and 0 were equally 292,132 records. Figure 4 shows an outcome of a 50-50 ratio, which satisfied the expectation of data resampling. Therefore, the dataset is balanced and ready for classification.

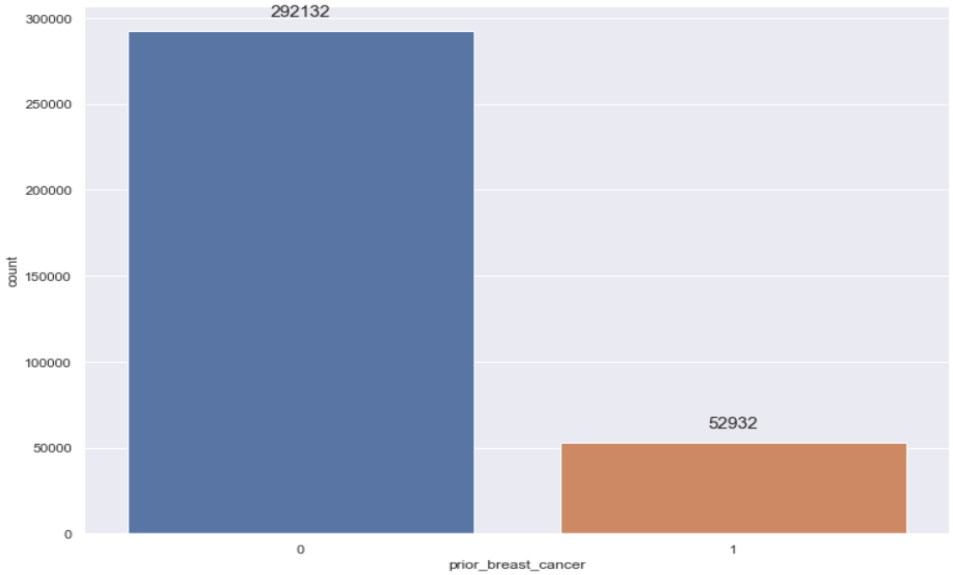

Figure 3: Count plot of imbalanced data

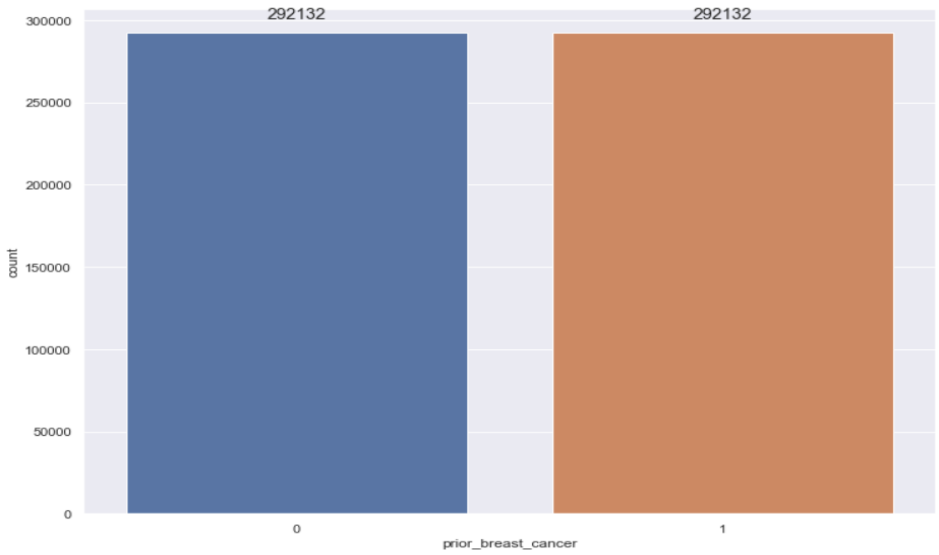

Fig. 4: Count plot after resampling (balanced data).

3.6. Classification

Four supervised machine learning techniques will be used in this research, which is Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). In terms of the classification techniques, 5-Fold Cross-Validation will be implemented with GridSearchCV. Cross-Validation divides the training set into k bins of equal size at random and each bin is trained with different learning experiments. On the other hand, GridSearchCV is a model selection under

the Scikit-learn package that enables the user to perform hyperparameter tuning to decide the ideal values for a model. This is a crucial process because it can act as a countermeasure to avoid any overfitting of the model.

During the classification, 3 train-test splits ratio are applied on the original dataset and the dataset with SMOTE. The data size for each train-test split set were shown in Table 3 and Table 4 respectively.

Table 3: Training and testing data size for original data.

Train-test Split	Training Set: Testing Set
80-20	276, 051: 69, 013
70-30	241, 544: 103, 520
60-40	207, 038: 138, 026

Table 4: Training and testing data size with SMOTE

Train-test Split	Training Set: Testing Set
80-20	467, 411: 116, 853
70-30	408, 984: 175, 280
60-40	350, 558: 233, 706

4.3. Evaluation metrics

Evaluating a model plays an important role in building an efficient machine learning model. This research carried out several evaluation measures such as accuracy, recall, precision, F1-score, and AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve to determine how well a model can perform.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

$$Precision\ = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1\ score\ =\ 2\ \times \left(\frac{Precision\ \times\ Recall}{Precision\ +\ Recall}\right) \quad (4)$$

Accuracy is a commonly used performance metric for evaluating the classifiers (Liu at al., 2014). It can be known as the percentage of correct classifications. Recall, also acknowledged as sensitivity is the percentage of true positive instances to the total actual positive instances in the data. Moreover, precision is the ratio of true

positive instances to all predicted positive instances by the model. It can signify how accurate the positive classification is.

When handling an imbalanced dataset, the class distribution is exceedingly skewed. In such a case, F1- score will be used to calculate the harmonic mean of precision and recall, where value of 1 is considered perfect while the worst is 0.

The ROC curve is a probability curve and the area below the curve is called AUC. True Positive Rate (TPR) is plotted against False Positive Rate (FPR) in the ROC curve. The AUC value of 1.0 represents a perfect prediction Narkhede et al. (2018). TPR and FPR are defined as follows.

$$TPR = \frac{True\ \ Positives}{True\ Postives\ \ + False\ Negatives} \tag{5}$$

$$FPR = \frac{False\ Positives}{False\ Positives\ \ + True\ Negatives} \tag{6}$$

5. Results And Discussions

5.1. Comparison results for Boruta

A preliminary result is performed to determine whether the features computed by Boruta should remain or be removed. First dataset without Boruta method was using the data which includes all the features. The second dataset removes the feature which obtains a Boruta score of 0.0. Another dataset will keep features with Boruta ranking above 0.3 only. All datasets are pre-trained and pre-tested on the LR classifier and each of them will undergo 3 different train-test splits. The comparative results are presented in Tables 5, 6 and 7.

Table 5: Comparison results for LR classifier taking all features.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8425	0.4526	0.0040	0.0078
70-30	0.8443	0.3333	0.0013	0.0026
60-40	0.8447	0.6250	0.0002	0.0005

Table 6: Comparison results for LR classifier with Boruta ranking above 0.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8423	0.3425	0.0023	0.0046
70-30	0.8446	0.5172	0.0009	0.0019
60-40	0.8447	0.5000	0.0001	0.0002

Table 7: Comparison results for LR classifier with Boruta ranking above 0.3.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8423	0.3019	0.0015	0.0029
70-30	0.8446	0.6667	0.0004	0.0007
60-40	0.8447	0.0000	0.0000	0.0000

Based on the preliminary results, there are not much difference in the accuracy with utilizing all features and removing some of the features. Since this dataset does not contains a lot of features, then this research work is taking all features for the remaining classification process.

5.2. Classification results for logistic regression (LR)

Tables 8 and 9 show the comparison classification results of LR for the data without SMOTE and with SMOTE respectively. By employing GridSearchCV for hyperparameter tuning, the best parameters chosen for LR are 0.1 for the value of C, l2 for penalty, and newton-cg for solver.

Table 8: LR comparison results without SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8425	0.4526	0.0040	0.0078
70-30	0.8443	0.3333	0.0013	0.0026
60-40	0.8447	0.6250	0.0002	0.0005

Table 9: LR comparison results with SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.7044	0.6974	0.7264	0.7116
70-30	0.7039	0.6971	0.7242	0.7104
60-40	0.7040	0.6972	0.7224	0.7096

5.3. Classification results for random forest (RF)

Tables 10 and 11 show the comparison classification results of RF for the data without SMOTE and with SMOTE respectively. By employing GridSearchCV for hyperparameter tuning, the best parameters for RF classifier in this research are None for max_depth, entropy for criterion. The value 5 for min_samples_leaf, 10 for min_samples_split. 150 for n_estimators was 150.

Table 10: RF comparison results without SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8426	0.0000	0.0000	0.0000
70-30	0.8446	0.7778	0.0004	0.0009
60-40	0.8447	0.0000	0.0000	0.0000

Table 11: RF comparison results with SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8196	0.7808	0.8906	0.8321
70-30	0.8151	0.7773	0.8847	0.8275
60-40	0.8094	0.7732	0.8763	0.8215

5.4. Classification results for support vector machine (SVM)

Tables 12 and 13 show the comparison classification results of SVM for the data without SMOTE and with SMOTE respectively. By employing GridSearchCV for hyperparameter tuning, the kernel is set as radial basis function (rbf), 0.1 for the value of gamma and 1 for C.

Table 12: SVM comparison results without SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8426	0.0000	0.0000	0.0000
70-30	0.8446	0.0000	0.0000	0.0000
60-40	0.8447	0.0000	0.0000	0.0000

Table 13: SVM comparison results with SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.7599	0.7250	0.8402	0.7784
70-30	0.7585	0.7239	0.8380	0.7768
60-40	0.7572	0.7231	0.8347	0.7749

5.5. Classification results for multilayer perceptron (MLP)

Tables 14 and 15 show the comparison classification results of MLP for the data without SMOTE and with SMOTE respectively. By employing GridSearchCV for hyperparameter tuning, the best parameters obtained for MLP are 0.0001 for alpha, relu for activation, adam for solver, (120, 80, 40) for the hidden layer size and adaptive for learning rate.

Table 14: MLP comparison results without SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.8426	0.0000	0.0000	0.0000
70-30	0.8446	0.0000	0.0000	0.0000
60-40	0.8447	0.0000	0.0000	0.0000

Table 15: MLP comparison results with SMOTE.

	Evaluation metrics			
Train-test	Test accuracy	Precision	Recall	F1-score
80-20	0.7133	0.6714	0.8402	0.7463
70-30	0.7349	0.6972	0.8332	0.7592
60-40	0.7271	0.7015	0.7918	0.7439

5.6. Classification model evaluation

All the classification processes are carried out on a computer with Intel(R) Core i5-8250U 1.80 GHz Processor, 12 GB RAM and additional Graphical Processing Unit (GPU), NVIDIA GeForce MX150.

For each set of experiment, LR took the shortest duration, which is about one hour, whereas RF took about 3 hours. MLP takes 20 hours, and SVM took 53 hours.

By comparing the model trained by using different train test splits, the results show that all the models trained by using data without resampling (without SMOTE) have a better test accuracy. However, the performances for the recall and F1-score were very low and even reach value of 0 for RF, MLP and SVM. This is due to the imbalanced of data which focus more on the majority class. Hence, this research is only considered the classification results from those experiments with involving SMOTE on the resampled data.

For overall model evaluation, 80-20 train test split performs the best among the other splits. This is because more data was used in training the model. Table 16 displays the classification results for all four models under 80-20 train test split.

Table 16: Evaluation results for classification models

	Accuracy	Precision	Recall	F1-score	AUC
LR	0.7044	0.6974	0.7264	0.7116	0.7732
RF	0.8200	0.7813	0.8905	0.8324	0.8712
SVM	0.7599	0.7250	0.8402	0.7784	0.8292
MLP	0.7081	0.6590	0.8669	0.7488	0.7828

RF performed the best among the four models with an accuracy score of 82%. RF is implemented by randomly taking subsamples to create decision trees and averaging the final decision to construct a refined model Wickramanayake (2020). When training the model, a tree is built from the subsample of training data. A series

of splitting will be performed, then the final decision is obtained by majority votes (Shichkin et al., 2018).

SVM scored higher accuracy than MLP, but the runtime for SVM was high (53 hours for SVM and 20 hours for MLP), this is mainly due to SVM translates n-dimensional spaces using kernel functions Thomas (2019). In MLP, sigmoidal activation function of neurons is employed as data signals propagate from the input layer, through the hidden layers and to the output layer. Despite direct processing hidden neurons, SVM optimized their parameters with kernel functions Osowski (2004).

Figure 5 displays the ROC curve for the comparison of models. The best accuracy of the test means that the curve is closer to the top-left corner; whereas if the curve is closer to the 45-degree diagonal line, it means that the model is less accurate for the test. In this research, the RF classifier, which is indicated in blue line shown closer to the top-left corner as compared to LR (red line), SVM (purple line), and MLP (green line).

Additionally, the RF obtained highest predictive accuracy, which can be seen from the Area Under the Curve (AUC). RF obtained the highest AUC value, which is near to value of 1.0. As the closer the value of AUC to 1.0, the better the model can predict. To conclude, RF has shown better performance than other classifiers in this research work.

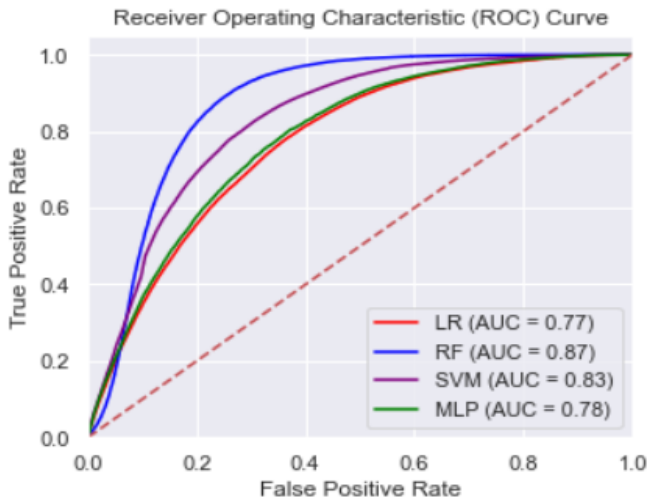

Fig. 5: ROC curve for the comparison models.

Table 17 shows the comparison of RF classification results and Kabir & Ludwig finding (2018). Although comparison with Kabir and Ludwig's work is unavoidable,

the testing of the models was performed on a more recent dataset which has a considerable number of additional samples.

Even though both datasets used in this research are obtained from the Breast Cancer Surveillance Consortium (BCSC) (BCSC, 2021), the dataset that used in this research consisted 1,522,340 records (collected between January 2005 and December 2017). The dataset used by Kabir & Ludwig (2018) consisted 1,144,565 records (collected between January 2000 to December 2009). Therefore, the year of collection might cause slight changes in the evaluation metrics results.

Table 17: Comparison with other approaches using same database and same RF classifier

	Proposed method	Kabir and Ludwig (2018)
No. of data	1,522,340	1,144,565
Year retrieve data	January 2005 to December 2017	January 2000 to December 2009
Classifier	Random Forest	
Accuracy	0.8200	0.8540
Precision	0.7813	0.9500
Recall	0.8905	0.8500
F1-score	0.8324	0.8900
AUC	0.8712	0.9140

6. Conclusion and Future Work

In conclusion, this research identified and classifies the risk factors of breast cancer by using four classifiers LR, RF, SVM, and MLP. These four classifiers undergo training and testing data with 80-20, 70-30, and 60-40 train test splits. The SMOTE data resampling technique shows significant improvements in the precision-recall rate. Besides, the recall score obtained from the proposed method outperforms Kabir's findings. The performances of each classifier were being recorded and the classification results was being compared. This research shows that RF performs the best performance with 82% of accuracy at 80-20 train test split.

For the future work, this research intends to extend by obtaining datasets from local hospitals and breast cancer organizations in Malaysia. Moreover, this research also plans to implement more classifiers to compare the performances different models.

Acknowledgments

The authors wish to express appreciation to National Cancer Institute, the Patient-Centered Outcomes Research Institute for publishing the dataset without charges and restriction.

References

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

Ballard-Barbash, R., Taplin, S. H., Yankaskas, B. C., Ernster, V. L., Rosenberg, R. D., Carney, P. A., & Kessler, L. G. (1997). breast cancer surveillance consortium: A national mammography screening and outcomes database. *American Journal of Roentgenology*, 169(4), 1001-1008.

Breast Cancer Surveillance Consortium (BCSC). https://www.bcsc-research.org/. (accessed Sep. 05, 2021).

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, 12(2), 119-126.

Kabir, M. F. & Ludwig, S. (2018, December). Classification of breast cancer risk factors using several resampling approaches. *In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 1243-1248).

Iqbal, H. N., Nassif, A. B., & Shahin, I. Classifications of breast cancer diagnosis using machine learning.

Lavanya, D. & Rani, K. U. (2013). A hybrid approach to improve classification with cascading of data mining tasks. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2(1).

Lehman, C. D., Arao, R. F., Sprague, B. L., Lee, J. M., Buist, D. S., Kerlikowske, K., & Miglioretti, D. L. (2017). National performance benchmarks for modern screening digital mammography: Update from the breast cancer surveillance consortium. *Radiology*, 283(1), 49.

Liu, Y., Zhou, Y., Wen, S., & Tang, C. (2014). A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, 6(4), 20-35.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards Data Science*, 26(1), 220-227.

Osowski, S., Siwek, K., & Markiewicz, T. (2004, June). MLP and SVM networks-a comparative study. *In Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004*, IEEE, 37-40.

Shajahaan, S. S., Shanthi, S., & ManoChitra, V. (2013). Application of data mining techniques to model breast cancer data. *International Journal of Emerging Technology and Advanced Engineering*, 3(11), 362-369.

Shichkin, A. V., Buevich, A. G., & Sergeev, A. P. (2018, July). Comparison of artificial neural network, random forest and random perceptron forest for forecasting the spatial impurity distribution. *In AIP Conference Proceedings*, AIP Publishing LLC , 1982(1), 020005.

Sickles, E. A., Miglioretti, D. L., Ballard-Barbash, R., Geller, B. M., Leung, J. W., Rosenberg, R. D., & Yankaskas, B. C. (2005). Performance benchmarks for diagnostic mammography. *Radiology*, 235(3), 775-790.

Thomas M. (2019). Comparing SVM and MLP machine learning models. Becoming human: Artificial Intelligence Magazine, 2019. https://becominghuman.ai/comparing-svm-and-mlp-machine-learning-models-348d08efea6b (accessed Mar. 31, 2022).

Wickramanayake B. (2020). random forest vs logistic regression. Medium. https://medium.com/@bemali_61284/random-forest-vs-logistic-regression-16c0c8e2484c (accessed Apr. 01, 2022).

Williams, K., Idowu, P. A., Balogun, J. A., & Oluwaranti, A. I. (2015). Breast cancer risk prediction using data mining classification techniques. *Transactions on Networks and Communications*, 3(2), 01.

World Health Organization, Breast cancer. (2021). https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Sep. 05, 2021).

Wolberg, W. H. & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences*, 87(23), 9193-9196.