EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Performance analysis of different machine learning algorithms in breast cancer predictions

Gopi Battineni*, Nalini Chintalapudi and Francesco Amenta

Telemedicine and Telepharmacy Center, School of Medicinal and Health Products Sciences, University of Camerino, Camerino, 62032, Italy.

Abstract

INTRODUCTION: There is a great percentage of failures in clinical trials of early detection of breast cancer. To do this, machine learning (ML) algorithms are useful to do diagnosis and prediction of cancer tumors with better accuracy.

OBJECTIVE: In this study, we develop an ML model coupled with limited features to produce high classification accuracy in tumor classification.

METHODS: We considered a dataset of 569 females diagnosed as 212 malignant and 357 benign types. For model development, three supervised ML algorithms namely support vector machines (SVM), logistic regression (LR), and K-nearest neighbors (KNN) were employed. Each model was further validated by 10-fold cross-validation and performance measures were defined to evaluate the model outcomes.

RESULTS: Both SVM and LR models generated 97.66% accuracy with total feature evaluation. With selective features, the SVM accuracy was improved by 98.25%. Whereas the LR model including limited features produced 100% of true positive predictions.

CONCLUSION: The proposed models involved by selective features could improve the prediction accuracy of a breast cancer diagnosis.

Keywords: Machine learning, feature selection, tumor classification, accuracy, AUC.

Received on 10 June 2020, accepted on 26 August 2020, published on 28 August 2020

doi: 10.4108/eai.28-5-2020.166010

*Corresponding author. Email: gopi.battineni@unicam.it

1. Introduction

Every year about 14 million people are suffering from different cancer types and most of them caused by the growth of malignant tumors [1]. Besides, breast cancer is among the second most common cancer type that identifies in women [2]. Statistics are saying that one in eight of the USA females were often exposed to breast cancers [3]. Experts were recommending mammography for early diagnosis of cancer to reduce the risk of mortality by 20-40% [4]. But by incorporating these methods, a high number of false negatives and false positives are kept remains the same, which limits the prediction accuracy. To avoid this, artificial intelligence (AI) techniques like machine learning are started to be involved for better cancer forecasting [5], [6].

Disease diagnosis by Machine learning (ML) algorithms is recently getting higher attention from communities of data science research [7]. It is because of the large adaption of computer-based techniques into different forms of healthcare and subsequent availability of medical databases [8], [9]. ML algorithms are associated with different probabilistic, statistical, and validation methods to learn from the experience and identify key data patterns from unstructured, large, and complex datasets.

This study primarily focuses on the performance analysis of malignant classification by different variants of ML algorithms. In the diagnosis of breast cancer, it is important to evaluate both false positives (i.e., no cancer but recommended for treatment) and false negatives (i.e., have cancer but not recommended for treatment). Therefore, ML algorithms are requested to identify and classify malignant (that has a high tendency of breast cancer) and benign (that has a low tendency of breast cancer) groups without bias.

Many studies were included in ML algorithms to do the early prediction of breast cancers. For example, cancer classification and prediction by logistic regression (LR) with Bayesian gene selection were effective to identify important genes with high accuracy [10]. It is reported that support vectors can predict cancer tumors by 89-97.5% of accuracy [6], [11], [12]. The idea behind cancer genes expansion was well explained through K-nearest neighbor with 92% accuracy [13]. In [14], researchers employed artificial neural networks to extract the malignant anomalies from fibrocystic breast masses. However, all these studies offered little explanations on the identification of true cancer positives and the importance of cell feature characteristic information to classify malignant tumors.

In this study, we highlight a feature dataset that is computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Single learning classifiers such as support vector machines (SVM), logistic regression (LR), and K-nearest neighbors (KNN) were considered on Wisconsin Breast Cancer Dataset (WBCD) to exhibit the high classification tasks. Moreover, the correlation matrix was derived to identify highly associated features in malignant classification. Model outcomes are further validated by other performance values like accuracy, sensitivity, the area under the receiver operating characteristic curve (AUROC or AUC).

2. Methods

2.1. WDBC dataset

We considered WBCD information, which obtained from the UCI machine-learning repository (https://archive.ics.uci.edu/ml/index.php). The dataset having 569 female information including 212 (37%) are diagnosed as the malignant type and 357 (63%) are diagnosed as benign type. Tumor or cancer cell features are extracted from a digital picture of fine-needle aspirates of breast masses. Dataset consists of 30 individual cell features including patient ID and diagnosis type. Ten real-valued features mentioned in Table 1 are computed for each cell nucleus and each image feature associated with three independent values such as mean, standard error (SE), and "worst" (mean of three largest values) resulting from 30 features.

In further, the correlation matrix was developed to identify high correlated features with tumor classification [15]. The correlation matrix is helpful to conduct data summary and understand the relationship between features and targeted outcomes. The heat map correlation matrix for the given features as depicted in Figure 1. We considered features with at least 80% of correlation with cancer diagnosis [16]. Seven features (i.e., radius_mean, perimeter_mean, area_mean, concave point_mean, radius_worst, perimeter_worst, and area_worst) were highly correlated to diagnosis outcome.

Table 1. Individual feature description of cell nucleolus

N	Feature	Description
1	Radius	The nucleus radius is defined as the distance between center to the points on the perimeter
2	Texture	The texture of the cell nucleus is measured by finding the standard deviation of gray-scale values
3	Perimeter	Perimeter calculated distance between the size of the core tumor
4	Area	Nucleus area is depending pixel count of snake interior, adding one half of the edge of the pixel
5	Smoothness	The smoothness of nuclear is measured by local variation in radius lengths
6	Compactness	The ratio of perimeter and area is producing the compactness, and is regulated by the formula: $Perimeter^2$ / are
7	Concavity	The severity of concave portions of the contour is defined as concavity
8	Concave Points	Number of concave portions of the contour
9	Symmetry	Measurement of length difference in between perpendicular lines to the central axis to the cell bounders in dual direction
10	Fractal Dimension	The fractal dimension was measured using "Coastline Approximation." Formula to approximate this value is Fractal dimension = Coastline approximation – 1

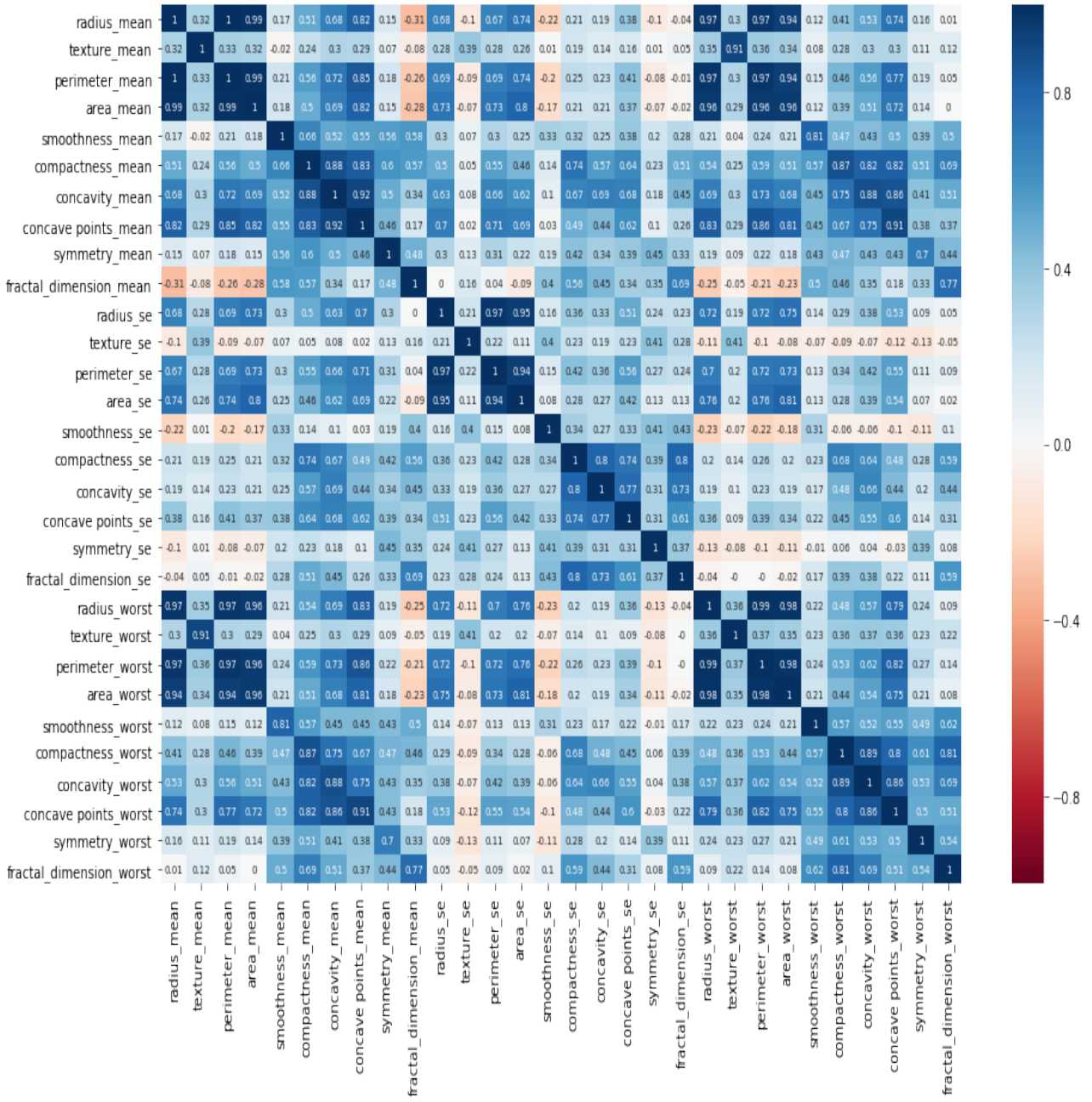

Figure 1. Correlation matrix heat map

2.2. Models

Support vector machines (SVM)

Support vector machines are famous machine learning algorithms to conduct outcome classification. The objective behind the SVM algorithms is to identify a hyperplane in N-dimensional space that randomly classifies the data points [17]. The hyperplanes are decision boundaries, which help to classify the data points. In two-dimensional space, the hyperplane is a line that separates data points into two segments. Moreover, the data points that can reside on either side of the hyperplane, which can be attributed to individual classes.

Logistic regression (LR)

Logistic regression is an effective binary classification algorithm and is used to define the datasets of discrete classes. Logistic functions offer a linear equation of

binary classification ranging from 0 to 1 [18]. Therefore, a binary LR model was considered during this study to classify cancer type. The simple logistic regression model can be defined by the equation.

$$\mathrm{Log}(n) = \frac{1}{1+e(-n)} \text{ where } 0 \leq n \leq 1$$

K-nearest neighbours (KNN)

The K-nearest neighbors are simple and basic ML algorithms, which are used for both regression and classification problems. These algorithms are popular as lazy learners because of passive nature in the resonance of large datasets. But these are highly useful to classify same-targeted outcomes.

2.3. Study framework

When the problem of diagnosing breast cancer occurs, it is important not to leave any true positives. Therefore, the highest sensitivity value of the ML algorithm can be selected. After identifying the features with high correlation, the data was splitting into two portions such as 70% (≈398 people) for train and 30% (≈171 people) for testing purposes. In further, model validation was conducted through k-fold cross-validation. In cross-validation, trained data was randomly partitioned into k folds of similar sizes, the k-1 folds are used for model training, and the rest one-fold was used for testing. The study framework is presented in Figure 2 and K=10 was considered for validation purposes.

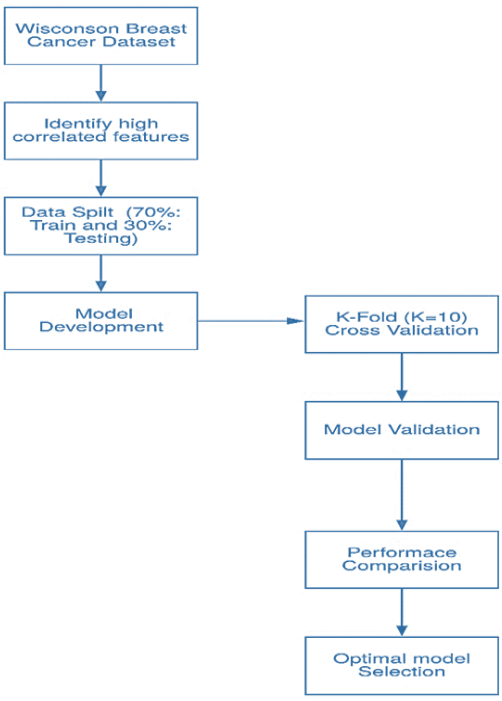

Figure 2. Study framework

2.4. Performance measures

During this study, we aim to identify or classify the malignant tumors from the WBCD. If the proposed ML algorithm left positive (M type) outcomes, then the patient could be at high risk of disease. To evaluate the maximum number of patients with a malignant tumor, we define performance measures like accuracy, sensitivity, precision, and AUC. Table 2 presents the confusion matrix example of the breast cancer dataset. Diagnosis classification has been done to identify malignant (severe cases) and benign (non-severe cases). We considered diagnosis is positive when X=M, and negative when Y=B.

Table 2. Confusion matrix of breast cancer masses

Classification	X	Y
X =M	TP	FN
Y =B	FP	TN

*TP: True positives, FP: False positives, FN: False negatives, TN: True negatives.

From the above confusion matrix, we define the following performance metrics.

Accuracy: percentage of total true predicted outcomes from total outcomes.

$$\text{Accuracy} = \left(\frac{TP+TN}{TP+TN+FP+FN} * 100\right)$$

Sensitivity: defines the proportion of true positives.

$$\text{Sensitivity} = \left(\frac{TP}{TP+FN} * 100\right)$$

Precision: Percentage of true positives from total positives.

$$\text{Precision} = \left(\frac{TP}{TP+FP} * 100\right)$$

3. Results

We can perform the model evaluation in different methods. This study not only aims to classify cancer diagnosis type but also to verify the adopted ML algorithm accurately classify the malignant cells. As discussed, the true positive rate (or sensitivity) addresses the question of how many female patients suffered from malignant tumors. It is also important to realize that high sensitivity does not guarantee high accuracy often there is also a tradeoff between individual performance measures.

We conduct two individual experiments: with total features and highly correlated features. Figures 3 presents a confusion matrix outcome three ML models of total

features and Figure 4 presents the confusion matrix outcomes three ML models of limited feature models.

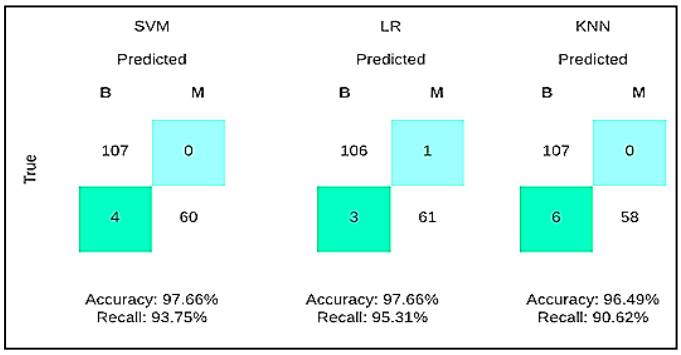

Figure 3. Confusion matrix outcomes of three models with total features.

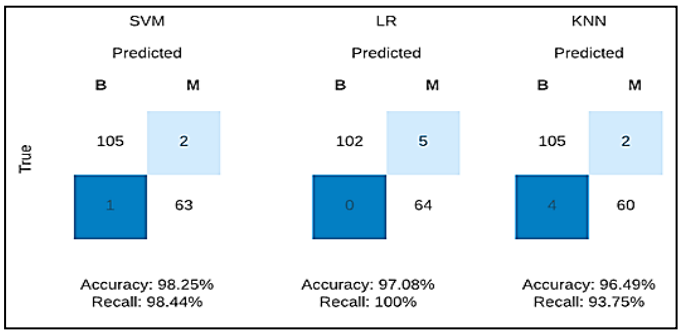

Figure 4. Confusion matrix outcomes of three models with selective features

Table 3 presents the performance comparisons of the two experiments on WBCD. In the first experiment, both SVM and LR got the highest accuracy (97.66%) which is followed by KNN by 96.49% of accuracy. However, LR possesses the highest sensitivity value of 95.31% and SVM, KNN has sensitivity 93.75%, 90.62% respectively. The precision of SVM, LR, and KNN was recorded as 100%, 98.38%, and 100% respectively. Also, we calculated the ROC to measure the correct classification of the three models. The ROC of SVM, LR, and KNN is generated as 0.9361, 0.9123, and 0.9371.

Table 3. Model outcomes (Total vs limited features)

Model	Accuracy	Sensitivity	Precision	ROC
SVM (Total)	0.9766	0.9375	1.0000	0.9361
SVM(Limited)	0.9825	0.9844	0.9692	0.9613
LR (Total)	0.9766	0.9531	0.9838	0.9123
LR(Limited)	0.9708	1.0000	0.9275	0.9913
KNN (Total)	0.9649	0.9062	1.0000	0.9371
KNN(Limited)	0.9649	0.9375	0.9677	0.9612

The first experiment produces better results but was comparatively low model performance when compared with existed studies. Therefore, the model evaluation was conducted by selective feature approaches. From Table 3, it is evident that diagnosis classification was done more accurately with limited features that were outperformed than total features. SVM generated the highest classification accuracy (98.25%), LR, and KNN recorded 97.08% and 96.49% of accuracy respectively. Simultaneously, LR classifies malignant tumors without bias and produced 100% of sensitivity value. It means that a better classification of malignant diagnosis was done. Moreover, SVM and KNN generated 98.44% and 93.75% of sensitivity. The precision of SVM, LR, and KNN was recorded as 96.92%, 92.75%, and 96.77% respectively. The AUC of SVM, LR, and KNN algorithms are recorded 0.9613, 0.9913, and 0.9612.

As mentioned in the last section, sometimes-left true positives are not affordable, and this could be fatal in cancer cases. Therefore, to find out cancers with malignant tumors, the present study highlights the importance of logistic regression with limited features. In further, AUC was assessed to present the relationship between the sensitivity (true positive rate) on the y-axis and specificity (false positive rate) on the x-axis. AUC associated with the LR model of selective features almost touches the highest value 1, which explains the ideal classification of cancer diagnosis (Figure 5).

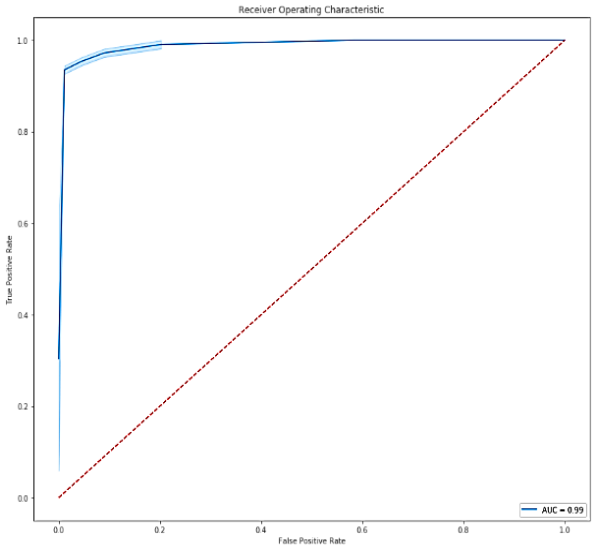

Figure 5. AUC curve LR model of selective features.

There are other similar established studies are available. The complex model using computer-aided diagnosis (CAD) systems were improved using a deep belief network and develop a complex classifier to test on WBCD [19]. However, the explanations offered by complex classifiers are still difficult for medical doctors to understand [20], but few studies are offered simple explanations by a single classifier approach. For example, the tumor binary classification was conducted and the highest accuracy (97.66%) was generated by SVM algorithms [21]. The comparative study of six ML

algorithms on WBCD reported that all models exceeded 90% test accuracy on the tumor classification task, and the multilayer perception algorithm remains standout one with 99% accuracy [22]. All the mentioned studies rather focused on the identification of true cancer positives (malignant), they mostly discussed ML classifier performances. Therefore, we aimed to propose the best ML classifier to identify malignant diagnosis with 100% accuracy.

4. Discussion

Currently, machine-learning models are being highly used in healthcare. It is mainly because compared to real-time clinical practices; ML models with a perfect algorithm can accurately diagnose any disease types [23]. ML algorithms in clinical practice can help doctors to identify medical data patterns and disease diagnosis from independent (trained) features [24]. However, the selection of a suitable machine learning algorithm is a key step because it depends upon data that used and targeted outcome [25]. These models can help the doctors in the early diagnosis of malignant tumors and ultimately save patients' life.

In this study, we considered the WBCD sample of 569 breast cancer diagnosed females. Thirty different features are computed from the digital image of the FNA tumor. In general, the classification accuracy of breast cancer tumors with fine-needle aspiration (FNA) is low when compared to ML techniques. Most of the established works adopt feature distinguish of cell nuclei to classify severe and non-severe cancer tumors. For instance, a study of FNA combined with matrix-assisted laser desorption utilizes principal component analysis to characterize lipid biomarkers and to define the accuracy of breast cancers [26]. An SVM based ensemble-learning algorithm was able to diagnose breast cancer with 97.89% of accuracy [27]. Although the mentioned studies were produced, better accuracy values no study highlights the classification of malignant tumors. In this paper, we developed an improved machine learning model with feature selection methods for breast cancer diagnosis. To do this we conduct two individual experiments to design models of total features and selective features. To ensure the correct identification of malignant tumors we highlight the sensitivity during the course of this study.

High correlated features associated with a cancer diagnosis were identified by the correlation matrix. The features with correlation ≥ 0.8 were further considered to do model development. Mean and worst feature values are largely associated with a cancer diagnosis. In the first experiment, the model was trained by total feature sets. Both SVM and LR models have classified diagnosis by 97.66% of accuracy. With selective features, SVM classified diagnosis by 98.25% and KNN by 96.49% of accuracies. However, while dealing with imbalanced datasets accuracy parameter not only defines the complete model performance. Therefore, other parameters like precision and sensitivity values are included to do further analysis.

As mentioned, this study mainly aims not to keep remain any malignant tumors. From Table 3, it is evident that logistic regression with limited features has perfectly classified the severe diagnosis patients. Ultimately, we validate the results with the AUC parameter and it will decide the maximum classification ranges. The AUC by LR algorithm with selective features was generated 0.99 that represents the ideal diagnostic accuracy of a cancer diagnosis.

By summing up, the presented results highlight that limited features selection of ML models will produce a great advantage of tumor classification over breast cancer studies. To our knowledge, the present study only produces 100% classification rate of malignant cells compared to other existing studies. However, used methods have some limitations in terms of applicability and training efficiency to other diseases limited by the dataset. We also recommend the method of identifying high correlated features in early diagnosis of other cancers like lung or prostate cancers. Finally introducing high correlation methods could accelerate the training pace for developed methods in terms of computation time.

5. Conclusion

In conclusion, this study presents the three improved single approached ML algorithms to identify high correlated features that are closely associated with malignant identification. The ML models proposed in this study can help to both health care professionals and medical researchers in breast cancer identification. We adopt three different supervised ML models namely SVM, LR, and KNN, and conduct the two different experiments (total features and limited features). Results have shown that three proposed models produce an accuracy 0.98, 0.97, 0.96 respectively and then the true positive (sensitivity) percentage of three algorithms are recorded as 98.44%, 100%, and 93.75% by selective features. However, the LR model with limited features could be the best solution to improve the diagnostic accuracy of breast FNA. Further studies are needed for confirming the study outcomes using large data of biomarkers, or multi-centric databases.

Conflicts of Interest

No author has any conflicts of interest

Acknowledgements.

Institutional funding of the University of Camerino supported Ph.D. bursaries to GB and NC.

References

[1] R. Siegel, E. Ward, O. Brawley, and A. Jemal, "Cancer Stastics, 2011," CA cancer J. Clin., 2011, doi: 10.3322/caac.20121.Available.

[2] Y. S. Sun et al., "Risk factors and preventions of breast cancer," International Journal of Biological Sciences. 2017, doi: 10.7150/ijbs.21635.

[3] I. A. for R. on Cancer, "Breast Cancer Estimated Incidence, Mortality and Prevalence Worldwide in 2012," 2012, 2012. .

[4] D. A. Berry et al., "Effect of screening and adjuvant therapy on mortality from breast cancer," N. Engl. J. Med., 2005, doi: 10.1056/NEJMoa050518.

[5] P. Folger, "Geospatial information and geographic information systems (GIS): Current issues and future challenges," in Geospatial Information and GIS: Background and Issues, 2011.

[6] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," 2016, doi: 10.1016/j.procs.2016.04.224.

[7] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," Journal of Personalized Medicine. 2020, doi: 10.3390/jpm10020021.

[8] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods," Machines, vol. 7, no. 4, p. 74, Dec. 2019, doi: 10.3390/machines7040074.

[9] B. Kaur, M. Sharma, M. Mittal, A. Verma, L. M. Goyal, and D. J. Hemanth, "An improved salient object detection algorithm combining background and foreground connectivity for brain image analysis," Comput. Electr. Eng., 2018, doi: 10.1016/j.compeleceng.2018.08.018.

[10] X. Zhou, K. Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," J. Biomed. Inform., 2004, doi: 10.1016/j.jbi.2004.07.009.

[11] W. Kim et al., "Recurrence Prediction Model for Breast Cancer," vol. 15, no. 2, pp. 230–238, 2012, doi: 10.4048/jbc.2012.15.2.230.

[12] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," Clin. Epidemiol. Glob. Heal., no. September, pp. 1–7, 2018, doi: 10.1016/j.cegh.2018.10.003.

[13] B. K. Singh, "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm," Biocybern. Biomed. Eng., vol. 39, no. 2, pp. 393–409, 2019, doi: 10.1016/j.bbe.2019.03.001.

[14] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn, and E. S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration," Cancer, vol. 116, no. 14, pp. 3310–3321, 2010, doi: 10.1002/cncr.25081.

[15] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," Stat. Comput., 2017, doi: 10.1007/s11222-016-9646-1.

[16] Y. Perez-Riverol, M. Kuhn, J. A. Vizcaíno, M. P. Hitz, and E. Audain, "Accurate and fast feature selection workflow for high-dimensional omics data," PLoS One, 2017, doi: 10.1371/journal.pone.0189875.

[17] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," pp. 1–13, 2006, doi: 10.11648/j.acm.s.2017060401.11.

[18] G. C. Cawley and N. L. C. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," Bioinformatics, 2006, doi: 10.1093/bioinformatics/btl386.

[19] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," Expert Syst. Appl., 2016, doi: 10.1016/j.eswa.2015.10.015.

[20] T. Miller, P. Hower, and L. Sonenberg, "Explainable AI: beware of inmates running the asylum," IJCAI 2017 Work. Explain. Artif. Intell., 2017, doi: 10.1016/j.foodchem.2017.11.091.

[21] J. Ivančaková, F. Babič, and P. Butka, "Comparison of different machine learning methods on Wisconsin dataset," 2018, doi: 10.1109/SAMI.2018.8324834.

[22] A. F. M. Agarap, "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," 2018, doi: 10.1145/3184066.3184080.

[23] G. Battineni, N. Chintalapudi, F. Amenta, and E. Traini, "A Comprehensive Machine-Learning Model Applied to Magnetic Resonance Imaging (MRI) to Predict Alzheimer's Disease (AD) in Older Subjects," J. Clin. Med., vol. 9, no. 7, p. 2146, Jul. 2020, doi: 10.3390/jcm9072146.

[24] M. Mittal, L. M. Goyal, S. Kaur, I. Kaur, A. Verma, and D. Jude Hemanth, "Deep learning based enhanced tumor segmentation approach for MR brain images," Appl. Soft Comput. J., 2019, doi: 10.1016/j.asoc.2019.02.036.

[25] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in Data Classification: Algorithms and Applications, 2014.

[26] Y. T. Cho et al., "Fine Needle Aspiration Combined With Matrix-assisted Laser Desorption Ionization Time-of-Flight/Mass Spectrometry to Characterize Lipid Biomarkers for Diagnosing Accuracy of Breast Cancer," Clin. Breast Cancer, 2017, doi: 10.1016/j.clbc.2017.04.014.

[27] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," Eur. J. Oper. Res., 2018, doi: 10.1016/j.ejor.2017.12.001.