Original Article Open Access

Archives Of Breast Cancer

DOI: 10.19187/abc.20207122-28

Accurate Detection of Breast Cancer Metastasis Using a Hybrid Model of Artificial Intelligence Algorithm

Jafar Abdollahi[a], Atlas Keshandehghan[b], Mahsa Gardaneh[c], Yasin Panahi*[a], Mossa Gardaneh[b]

[a] *Deputy of Research and Technology, Ardabil University of Medical Sciences, Ardabil, Iran*
[b] *Department of Stem Cells and Regenerative Medicine, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran*
[c] *Department of Biology, York University, Toronto, Canada*

ARTICLE INFO

Received:
15 December 2019
Revised:
08 January 2020
Accepted:
18 January 2020

Key words:
Breast cancer, metastasis, machine learning, specificity selection, classification algorithms, hybrid algorithm.

ABSTRACT

Background: Breast cancer (BC) is a prevalent disease and a major cause of mortality among women worldwide. A substantial number of BC patients experience metastasis which in turn leads to treatment failure and death. The survival rate has been significantly increased due to more rapid detection and substantial improvements in adjuvant therapies including newer chemotherapeutic and targeted agents, and better radiotherapy techniques.

Methods: In this study, we cross-compared the application of advanced artificial intelligence algorithms such as Logistic Regression, K-Nearest Neighbors, Discrete Cosine Transform, Random Forest Classifier, Support Vector Machines, Multilayer Perceptron, and Ensemble to diagnose BC metastasis. We further combined MLP with genetic algorithm (GA) as a hybrid method of intelligent analysis. The core data we used for comparison belonged to the images of both benign and malignant tumors collected from Wisconsin Breast Cancer dataset from the UCI repository.

Results: The application of several different algorithms to the collection of BC data indicated that these algorithms have comparable accuracy rate in detecting and predicting cancer. However, our hybrid algorithm showed superior accuracy, sensitivity and specificity compared to the individual algorithms. Two methods of comparison (Cross-Validation and Holdout) were applied to this study which produced consistent results.

Conclusion: Our findings indicate that our MLP-GA hybrid algorithm can speed up diagnosis with higher accuracy rate than the individual patterns of algorithm.

Introduction

Breast cancer (BC) is one of the most prevalent cancer types among human populations with one out of eight women being reportedly affected during her lifetime. The increase of BC outbreak has been significant in the last fifty years causing the highest number of deaths among all cancer types. In 2019 alone, approximately 268,600 women were expected to be diagnosed with invasive BC with 62,930 women within situ BC in the United States.[1] The mortality rate due to BC is estimated to be 42,260 deaths (41,760 women and 500 men). Sixty-two percent of BC patients are diagnosed with a tumor localized in the breast whose 5-year survival rate is 99%.[2] However, the rate is reduced to 85% for those who have the cancer metastasized to the regional lymph nodes and it will be dramatically dropped to 27% if the cancer spreads to distant organs.[3,4]

Although most cancers originate from a mutated single cell, tumors grow and gain heterogeneous cell populations, so they have no common genotypic/phenotypic profile. This heterogeneity is translated

***Address for correspondence:**
Yasin Panahi, PhD
Address: Pajoohesh Blv, Tehran-Karaj HWY, Tehran, 14965/161, Iran
Tel: +98 21 44580344
Fax: +98 21 44580395
Email: y.panahi@arums.ac.ir, panahi_yasin@yahoo.com

to some cancer cell populations with invasive and metastasizing capacity causing tumor malignancy.[5] Metastasis is a biological hallmark of malignant tumors and the main cause of cancer mortality.[6] Therefore, early detection can have extremely profound impact on overall patient outcome and survival.

BC is a systemic disease that may be treated via surgery and chemo- and radiotherapy.[7] Pathologically, a single factor alone cannot be attributed to BC, scince there are several factors that are proposed as BC risk factors including age, personal history, breast pathology, family history, and genetic predisposition.[8] Also, interactions between environmental and genetic factors can induce susceptible genes such as BRCA1 and BRCA2 to undergo mutation,[9] which in turn leads to tumorigenesis.

Current diagnostic methods of tumor metastasis are developed as preclinical and clinical assays. Preclinical assays are used in ex-vivo samples and in in-vivo mice and comprise lab tests/histopathology, non-invasive blood tests, small animal imaging, molecular genetic imaging, and circulating tumor cells. Approaches used to detect metastatic lesions at clinical levels include biomarker tracking, imaging procedures, and circulating tumor cells.[10] In fact, a diagnostic test is considered complete when it can achieve 100% sensitivity and specificity. [11]None of these diagnostic methods meet this criterion and have not yet been successfully translatable from bench to bedside.

Some of these methods have been adopted for cancer diagnosis for more than 40 years. However, the biopsy-based histopathology is one of the most widely used methods to detect metastatic cells and classify cancer. Pathologists use histopathological images to accurately identify tumor cells. With this technique, they examine the microscopic structure of the patient's tissue.[12] Of note, this is the gold standard for the detection and diagnosis of micrometastases in tissues suspected to have tumor cells.[13] However, analysis of slides made from biopsies is a laborious task for a pathologist and requires sufficient time, remarkable skill and prudence. It is not an error-free task and might leave small metastases un-detected.[14] In comparison, computer-aided histopathological analysis can play an important role in the detection and prediction of BC tumor cells.[15] In this case, the use of Artificial Intelligence (AI) algorithms is an attempt to improve speed and accuracy of diagnosis.

Several AI algorithms have been developed and applied to cancer detection. Logistic Regression (LR) is a statistical and mathematical procedure that explores data sets with one or more free variants that hold a result, whereas K-Nearest Neighbors (KNN) is a supervised machine learning method used to solve classification and regression problems. Decision Tree Classifier (DTC) works by separating images into parts of differing importance and converts them into equivalent frequency coefficients. Random Forest Classifier (RFC) is another algorithm based on an ensemble learning method that creates a set of decision trees from randomly-selected subset of training set to aggregate votes and decide the final class of the test object. Support Vector Machines (SVM) is a supervised learning method that examines the data and sorts it into one of the two categories. Multilayer Perceptron (MLP), on the other hand, is a feed-forward artificial neural network that generates a set of outputs from a set of inputs. Finally, Ensemble techniques are machine learning techniques where more than one learner are constructed for a given task, and Multilayer Perceptron/Genetic Algorithm (MLP-GA) can more accurately detect tumor cells. Given that the individual use of AI algorithms may not lead to rapid and accurate detection of tumor cells, we attempted to cross-compare a number of these index algorithms in our study. We also combined MLP with genetic algorithm (GA) to create a hybrid pattern for boosted efficiency of detection.

Methods

In this study, we used data of digitalized pathologic images from Wisconsin Breast Cancer dataset [(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)] that included 357 benign and 212 malignant BC samples. We primarily reviewed the reports of the period 2006-2019 on diagnosis and classification of images collected from cancer tissues based on various algorithms. Our review indicated that application of an ML algorithm alone has not been precise and successful in detection and prediction of diseases. Next, we used 7 standardized and widely used algorithms to classify BC on our samples for comparison. K-Nearest Neighbors (KNNs) that uses the probability of density estimation based on nuclear methods, and Decision Tree (DT) employed for pattern extraction were the first two algorithms we applied. Our third algorithm was Forest Tree (FT) and we used a collection of FTs to classify cancer patients. We entered a set of data to each FT so the algorithm could start learning. For prediction, we used a new set of data so the FTs could predict the outcome.

Logistic Regression (LR) was the fourth algorithm employed in order to segregate benign tumors from malignant ones. For this, we applied 30 parameters as potential risk factors and compared the results using Receiver Operating Characteristics (ROP) in terms of the level of sensitivity and specificity. In MLP that simulates human brain performance[16], we used Gradient-based Multilayer Perceptron neural network as an MLP type widely applied to BC diagnosis. The next algorithm we used was SVC(Support Vector Classifier) the kernel of which was used for the classification of polished data

and enhanced precision of benign/malignant tumor diagnosis. After the basic algorithms, group learning algorithm was applied to develop a model of data classification with enhanced precision and performance compared to the basic algorithms.

Finally, in order to boost neural network algorithms such as MLP, we applied genetic algorithms (GAs) to add a new hybrid pattern of MLG-GA to individual algorithms. Indeed, we used GA to optimize initial weights in a neural network and determine optimal quantities of the parameters in both learning and detection processes of the system. We adopted two different strategies called Cross-Validation and Holdout to our comparison studies. We applied 569 samples in each experiment and devoted 60% for learning, 20% for evaluation, and 20% for the main tests. For this, we applied various criteria that included sensitivity, accuracy and specificity, True/False Positive Rates, Positive/Negative Predictive Values, methods of data segregation, etc.

Results

Algorithms: an overview

Figure 1 shows a flow chart of the algorithms we applied in this study. It shows how our hybrid system is designed. We entered the data into our machine learning algorithms in 3 steps. In step one, basic algorithm was applied individually but in step 2 they were then used in groups. In step 3, we applied our hybrid system comprising of MLP and genetic algorithm (MLP-GA). The algorithms were ultimately evaluated for their performance in predicting the samples.

Comparison of algorithms for their accuracy of diagnosis

Table 1 compares the performance of different algorithms using the Holdout strategy. Data achieved through KNN indicated that Gaussian Nuclear Density Estimation based on Euclidean distance detects cancer with 93% accuracy.

Data collected from DT and FT showed, respectively, 93% and 97% accuracy. Logistic Regression predicted cancer with 96% accuracy, 94% sensitivity and 95% specificity whereas MLP produced 95% accuracy. The accuracy figures for SVM and group learning were found, respectively, 96% and 93%. Figure 2 compares the accuracy rates based on the Holdout approach.

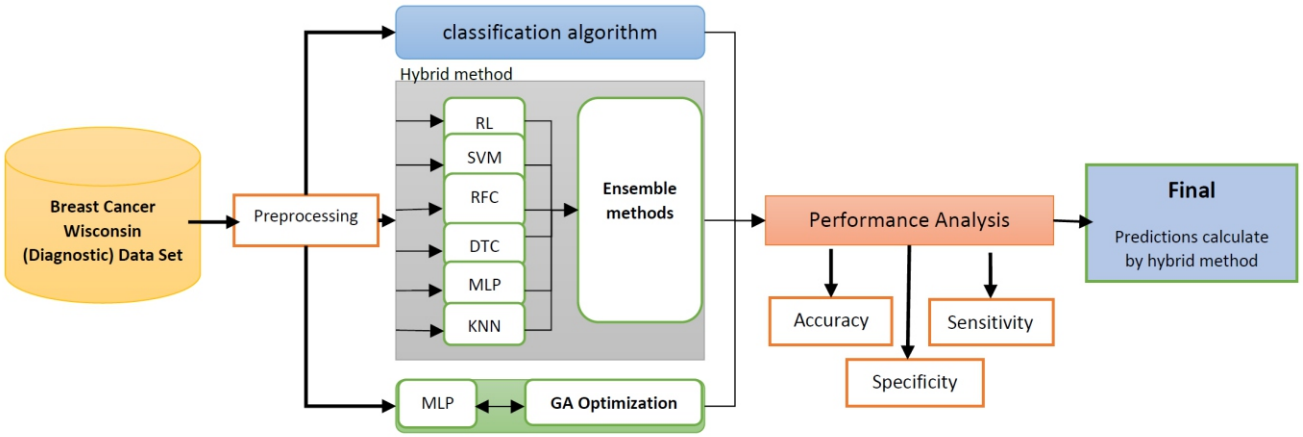

Figure 1. A schematic flow chart of the algorithms.

Table 1. Comparison of accuracy rates.

Algorithm	Accuracy	Standard deviation	Precision	Recall	F1-score
LR	0.96	0.00987	0.96	0.97	0.96
SVM	0.96	0.00987	0.96	0.97	0.96
RFC	0.97	0.00970	0.97	0.97	0.97
DTC	0.93	0.01101	0.94	0.94	0.94
MLP	0.95	0.00943	0.95	0.95	0.95
KNN	0.93	0.00966	0.94	0.94	0.94
Ensemble	0.93	0.00966	0.94	0.94	0.94
MLP-GA	0.98	0.00543	0.98	0.98	0.98

Diagnosis of BC was compared in rates among various algorithms with Holdout approach.

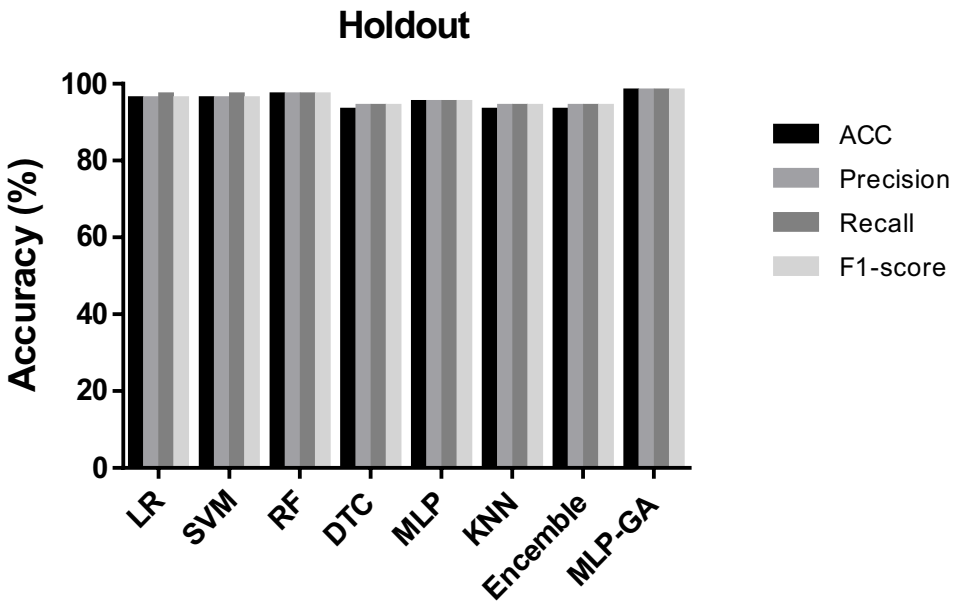

Figure 2. Comparison of accuracy rates. Graph shows the percentage of accuracy for each algorithm using Holdout comparison approach. Each column represents at least two different comparison experiments carried out independently.

Comparison between these algorithms by the Cross-Validation strategy generated data that was nearly identical to those of the Holdout approach. This showed that the MLP-GA hybrid has the highest accuracy rate among all those algorithms used in our study (Table 2 and Figure 3).

We employed boxplot in order to demonstrate the accuracy rate of our algorithms in classifying cancer data using cross-Validation. Figure 4 shows the outcome of the boxplot analysis in the distribution of accuracy rates for the validation of the algorithms. A boxplot was drawn for each figure of diagnosis algorithms. The results indicate that all algorithms have a range of the least distance between 86 and 100 with an average of 94.

Table 2. Comparison of accuracy rates.

Algorithm	Accuracy	SD
LR	0.93	0.01839
KNN	0.93	0.01282
DTC	0.92	0.01121
RFC	0.95	0.01204
SVM	0.93	0.01966
MLP	0.96	0.00966
Ensemble	0.94	0.01615
MLP-GA	0.997	0.00015

Diagnosis of BC was compared in rates among various algorithms with Cross-Validation approach. LR, Logistic Regression; KNN, K-Nearest Neighbors; DTC, Decision Tree Classifier; RFC, Random Forest Classifier; SVM, Support Vector Machines; MLP, Multilayer Perceptron; MLP-GA, Multilayer Perceptron/ Genetic Algorithm; SD, Standard Deviation.

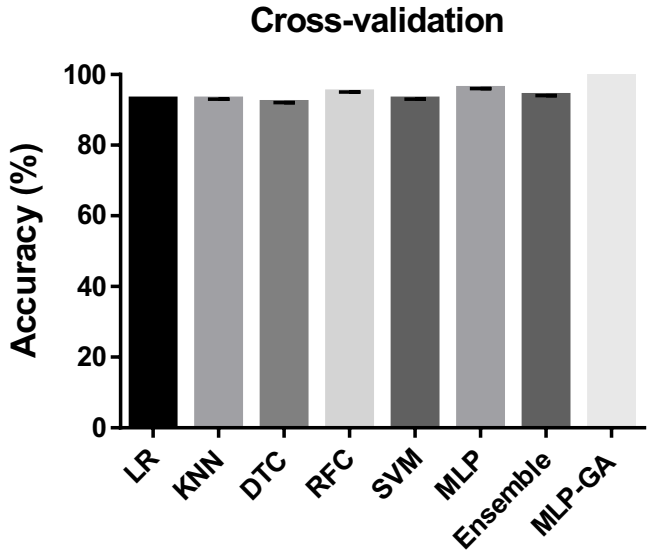

Figure 3. Comparison of accuracy rates. The graph shows the accuracy percentage for each algorithm using cross-validation comparison approach. Each column represents at least two different comparison experiments carried out independently.

We next applied randomized search (RS) and grid search (GS) methods in order to optimize estimating parameters and improve our results as well as find those parameters that affect the learning process in parallel. Figure 5 shows the result of comparison and evaluation for each algorithm using RS and GS. These findings indicate that our hybrid algorithm MLP_GA has a superior performance in optimization and search for parameters effective in rapid diagnosis of cancer with an accuracy rate of 98%.

In this study, we used various algorithms for classifying BC datasets. In order to select the best algorithm, we applied the timing criterion and its complexity as an element for comparing multiple algorithms for problem solving. Table 3 shows the results of comparing execution time complexity within various algorithms. As shown in Table 3, the MLP-GA algorithm has higher performance in terms of time complexity and in classifying datasets. Its execution time reached 0.04 with an accuracy rate of 98.3%.

We, therefore, used Confusion Matrix (CM) to show the overall performance of algorithms so we could evaluate their function or malfunction in the classification and diagnosis of datasets. Table 4 compares the performance of various algorithms using CM.

The data indicate the high performance of our recommended algorithm MLP-GA so that its accuracy, sensitivity and specificity have reached, respectively, 98%, 98% and 94%. They also indicate that the algorithm MLP_GA has a superior performance and is capable of classifying all datasets.

Hybrid algorithm MLP-GA with the highest level of accuracy and lowest deviation

We showed that our hybrid system has an accuracy rate of 99.7% which was the highest among all algorithms we studied (Table 2 and Figure 3). This system also showed a deviation rate of 0.00015 disease prediction, the lowest rate compared to those of other algorithms we studied.

Discussion

Breast cancer is a malignant tumor with different phenotypes. We applied AI-based effective algorithms to breast tissue images and addressed the question of whether AI can help detect cancerous tissues and BC metastasis as a result. We particularly found that a

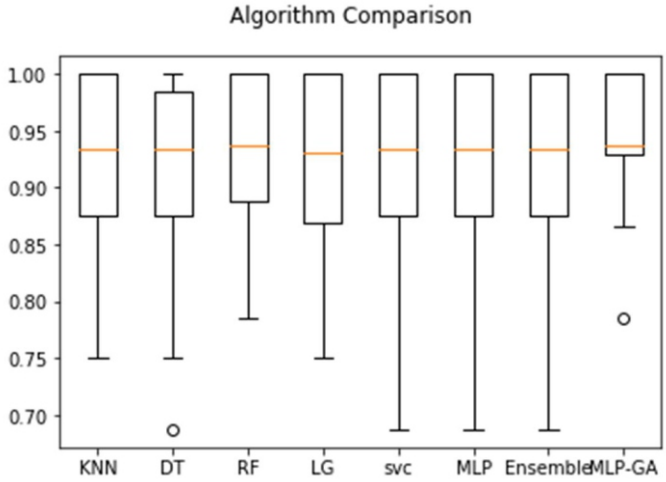

Figure 4. Comparison of accuracy rates among algorithms in the distribution of accuracy rates in validating classification data.

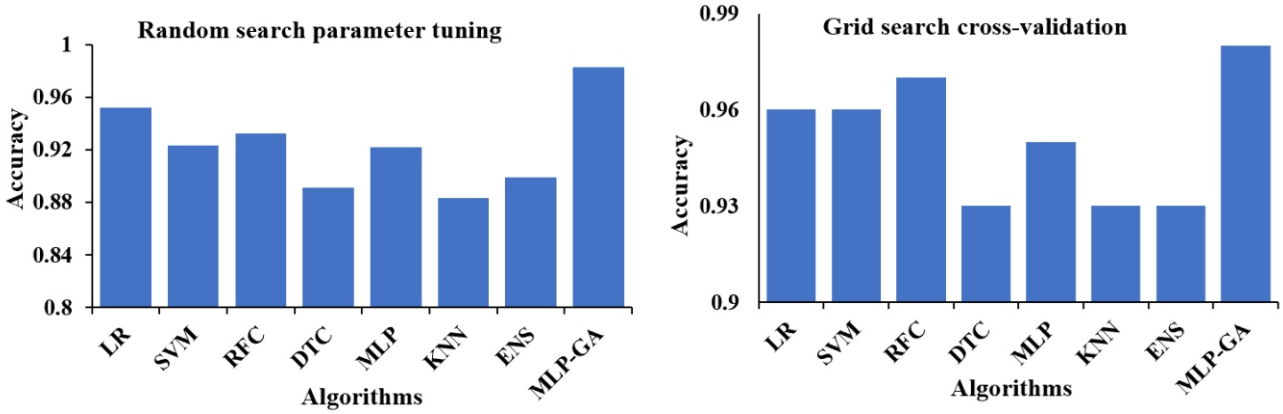

Figure 5. Comparison of accuracy rate of algorithms for optimization and search for parameters effective in cancer detection

Table 3. Comparison of timing complexity in the classification of datasets.

Algorithm	Accuracy	SD	Precision	Recall	F1-score	
KNN	0.120921	96	0.93	0.95	0.96	0.95
DT	0.099941	96	0.93	0.92	0.96	0.92
RF	0.442738	97	0.92	0.92	0.97	0.93
RG	0.101936	93	0.95	0.89	0.93	0.89
SVM	0.117928	95	0.93	0.92	0.95	0.92
MLP	0.124924	93	0.96	0.88	0.93	0.88
Ensemble	0.112928	93	0.94	0.89	0.93	0.89
MLP-GA	0.046721	98	0.997	0.98	0.983	0.983

Table 4. Comparison of performance among various algorithms using Confusion Matrix.

Algorithm	LR	SVM	RFC	DTC	MLP	KNN	Ensemble	MLP-GA
Accuracy	l0.96	0.96	0.97	0.93	0.95	0.93	0.93	0.98
Sensitivity	0.94	0.95	0.96	0.93	0.95	0.93	0.91	0.98
Specificity	0.95	0.95	0.97	0.93	0.95	0.89	0.89	0.94
PPV	0.90	0.90	0.88	0.89	0.91	0.90	0.86	0.98
NPV	0.90	0.89	0.90	0.90	0.90	0.89	0.86	0.98
TPR	0.92	0.90	0.96	0.90	0.89	0.90	0.88	0.98
FPR	0	0	0	0	0	0	0	0

combined algorithmic procedure of MLP-GA can more efficiently minimize errors of diagnosis.

Of the AI algorithms that we examined, ANN identifies patterns and processes that originate from the biological neural system and operates similar to the brain. ANN is comprised of a vast number of extraordinary interconnected processing elements called neurons that act in concert to solve a problem. It has been applied to BC detection[16] and risk assessment.[18, 19] K-Nearest neighbors algorithm is a non-parametric method used for classification and regression. This approach has also been applied to BC detection.[20, 21] Random Forest or Random Decision Forest is considered a supervised learning algorithm that builds a forest by accident. Accidental Forest makes several Decision Trees and then links them to produce more accurate and reliable predictions. It has been applied to BC detection.[22]

All the algorithms we used to detect BC worked with reasonable accuracy and speed. However, they differ from one another in their composition and functionality. In order to benefit from their strength and enhance the efficiency of diagnosis, we created the MLP-GA hybrid algorithm that demonstrated maximum precision with minimum errors. Similar attempts have been already made to detect heart disease and Coronary Artery Disease.[23-25] In an elegant study, Mobadersany et al. applied their convolutional neural networks to combine histology images and genomic biomarkers into a single unified framework and showed that this hybridized system is superior to current procedures in predicting the overall survival of glioma patients.[26] Belciug and Gorunescu applied a hybrid neural network/genetic algorithm to detect BC occurrence and recurrence. They designed a multi-layer perceptron using a GA routine and found that this combination is superior to individual approaches in providing accurate classification of BC samples.[27]

Our hybrid system, in line with the systems reported above, recommends precise therapeutic procedures for complex diseases, reduces errors in clinics and improves enrolment in clinical trials. Since AI technology is rapidly growing in medicine, it will practically address data sharing and privacy, the transparency of algorithms, standardization of data and cooperation in various operating systems. Our main objective in the near future is designing intelligent models for BC metastasis and image classification assisted by early diagnosis of cancer using AI particularly MLP.

In conclusion, our hybrid algorithm MLP-GA was found superior to our other algorithms in detecting BC incidence more reliably and accurately.

Conflict of Interests

None.

References

1. American Cancer Society: https://www. cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf.
2. Cancer Facts & Figures 2018. Atlanta: American Cancer Society; 2018.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: A Cancer J Clin. 2018; 68(1):7-30.
4. Breast Cancer Facts & Figures 2017-2018.

Atlanta: American Cancer Society, Inc.; 2017.
5. Meacham C, Morrison S. Tumour heterogeneity and cancer cell plasticity. Nature. 2013. 501:328–337.
6. National Cancer Institute, 2019, https://www.cancer.gov/types/metastatic-cancer.
7. Sarıtaş AG, Yalav O, Kekeç Y, Sakman G, *et al.* Effect of neoadjuvant chemotherapy on estrogen receptor, progesterone receptor, Cerb-B2, vascular endothelial growth factor and Ki-67 in patients with locally advanced breast cancer. Cukurova Med J. 2019; 44(1):226-231.
8. Shah R, Rosso K, Nathanson SD. Pathogenesis, prevention, diagnosis and treatment of breast cancer. World J Clin Oncol. 2014;5(3):283–298.
9. Nathanson, KN, Wooster R, & Weber BL. Breast cancer genetics: what we know and what we need. Nature medicine, 2001; 7(5):552.
10. Menezes ME, Das SK, Minn I, Emdad L, Wang XY, Sarkar D, Pomper MG, Fisher PB. Detecting Tumor Metastases: The Road to Therapy Starts Here. Adv Cancer Res. 2016; 132: 1-44.
11. Mordente, A., Meucci, E., Martorana, G. E., & Silvestrini, A. Cancer Biomarkers Discovery and Validation: State of the Art, Problems and Future Perspectives. Adv Exp Med Biol. 2015; 867:9-26.
12. Kumar R, Srivastava R, & Srivastava S. Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features. J Med Eng. 2015:14.
13. Bird B, Romeo M, Laver N, & Diem M. Spectral detection of micro-metastases in lymph node histo-pathology. J Biophotonics. 2009; 2(1-2):37-46.
14. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A Dataset for Breast Cancer Histopathological Image Classification. IEEE Trans Biomed Eng. 2016; 63:1455-1462.
15. Aswathy M, & Jagannath M. Detection of breast cancer on digital histopathology images: Present status and future possibilities. Informatics in Medicine Unlocked. 2017; 8:74-79.
16. Abdollahi J, Moghaddam BN, & Parvar ME. Improving diabetes diagnosis in smart health using genetic-based Ensemble learning algorithm. Approach to IoT Infrastructure. Future Gen Distrib Systems J. 2019; 1:23-30.
17. Saritas I. Prediction of Breast Cancer Using Artificial Neural Networks. J Med Syst (2012) 36: 2901.
18. Sepandi M, Taghdir M, Rezaianzadeh A, Rahimikazerooni S. Assessing Breast Cancer Risk with an Artificial Neural Network. Asian Pac J Cancer Prev. 2018; 19(4):1017-1019.
19. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE Jr, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer. 2010;116(14):3310–3321.
20. Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression. The Am Stat. 1992; 46 (3): 175–185.
21. Sarkar M, Leong TY. Application of K-nearest neighbor's algorithm on breast cancer diagnosis problem. Proc AMIA Symp. 2000;759–763.
22. Zhang H, Wang M. Search for the smallest random forest. Stat Interface. 2009; 2(3):381.
23. Liu X, Wang X, Su Q, *et al.* A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Comput Math Methods Med. 2017; 2017:8272091.
24. Verma L, Srivastava S, & Negi PC. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. J Med Syst. 2016, 40, 178 (2016).
25. Kolukisa B, Hacilar H, Goy G, Kus M, Bakir-Gungor B, Aral A, Gungor VC. Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. 2018 IEEE International Conference on Big Data: 2232-2238.
26. Mobadersany P, Yousefi S, Amgad M, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018;115(13):E2970–E2979.
27. Belciug S, Gorunescu F. A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence. Expert Systems. 2013; 30(3):243-254.