TPM Vol. 32, No. S2, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

MACHINE LEARNING IN PERSONALIZED MEDICINE FOR BREAST CANCER DIAGNOSIS

[1]D. LEELA DHARANI, [2]MOHAMMED ALI SOHAIL, [3]TASNEEM BANO REHMAN, [4]M JAYAPRAKASH, [5]C. KALPANA, [6]BANSODE G.S

[1]ASSISTANT PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY, P.V.P SIDDHARTHA INSTITUTE OF TECHNOLOGY, VIJAYAWADA, ANDHRA PRADESH, INDIA. DHARANIDONEPUDI@GMAIL.COM
[2]LECTURER, DEPARTMENT OF ELECTRICAL & ELECTRONIC ENGINEERING, COLLEGE OF ENGINEERING & COMPUTER SCIENCE, JAZAN UNIVERSITY, JAZAN, K.S.A. MSOHAIL@JAZANU.EDU.SA
[3]ASSOCIATE PROFESSOR, COMPUTER SCIENCE AND ENGINEERING DEPARTMENT, MUFFAKHAM JAH COLLEGE OF ENGINEERING &TECHNOLOGY, HYDERABAD, INDIA. TASNEEM.BANO@GMAIL.COM
[4]PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SAVEETHA SCHOOL OF ENGINEERING, SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES, SAVEETHA UNIVERSITY, CHENNAI, TAMILNADU, INDIA. JAYAPRAKASH040387@GMAIL.COM
[5]ASSISTANT PROFESSOR, INFORMATION TECHNOLOGY, KARPAGAM INSTITUTE OF TECHNOLOGY, COIMBATORE, TAMILNADU, INDIA. KALPANA.IT@KARPAGAMTECH.AC.IN
[6]ASSISTANT PROFESSOR, DEPARTMENT OF ENGLISH, KONERU LAKSHMIAH EDUCATION FOUNDATION, KL (DEEMED TO BE) UNIVERSITY, VIJAYAWADA-GUNTURU, ANDHRA PRADESH, INDIA. BANSODEGS@KLUNIVERSITY.IN

Abstract:
Aim: The aim of this work is to to improve the prognostic precision in breast cancer by integrating machine learning with proposed methods..
Background: In personalized medicine, the utilization of learning approaches has become increasingly significant as a result of evaluation of complex, large-scale, and unstructured data and information. In recent years, a significant number of researchers have shown an interest in personalized medicine, which entails the development of one-of-a-kind treatments for each individual patient on the basis of their shared traits, which may include their DNA, their heredity, and their way of life.
Problem: a In addition to being one of the malignancies, breast cancer is characterized by a wide range of subtypes that exhibit a diversity of clinical outcomes. Breast cancer has a number of diverse biological origins. As a consequence of this, proper disease stratification is very necessary in order to provide individually tailored therapeutic therapy for breast cancer.
Methodology: For the purpose of developing prognostic models for the advancement of breast cancer, this study utilized machine learning in conjunction with random optimization (RO) to incorporate glucose metabolism markers and other prognostic characteristics into a dataset. There were many different performance metrics that were utilized in order to evaluate the models.
Results: There was a high level of analytical performance among the ML-proposed models, with AUC values of 0.75 or above; among these models, the ML-RO-0 model had the greatest relative significance for glucose metabolism features. It was determined that the combined DSS model was successful with a c-statistic of 0.84.

Keywords: Breast cancer, machine learning, random optimization, prognostic models, personalized medicine.

INTRODUCTION

Relatively few studies have been conducted to determine whether or not they are applicable to cancer prognosis (Tran et al., 2019; Ming et al., 2019; Ahn et al., 2023; Vadapalli et al., 2022; Low et al., 2018). HER2/Neu expression has been recognized as having a significant prognostic relevance; however, this information was not included in the SEER dataset that was utilized in this particular instance (Alzu'bi et al., 2021). As a result, it is abundantly evident that there is a requirement for the development of prognostic categorization models that combine the most recent advancements in artificial intelligence technology (Sugimoto et al., 2023).

Related Works

According to the National Institutes of Health, "a clinical useful prognostic biomarker must be a proven independent (Rezayi et al., 2022), significant factor that is easy to determine and interpret and that has therapeutic consequences" (Manikis et al., 2023; Sotudian & Paschalidis, 2021). A predictive biomarker can tell us how a patient's cancer will ultimately turn out, regardless of the therapeutic response that the patient receives (Saravanan et al., 2023). It is for this reason that prognostic biomarkers, despite the fact that they can be utilized to select individuals who would receive adjuvant systemic treatment, are unable to accurately predict how effectively the treatment will be administered (Yuvaraj et al., 2022).

If prognostic biomarkers were able to more accurately signal the likelihood of recurrence, then a sizeable percentage of patients would be able to avoid the potentially damaging consequences of chemotherapy without compromising their chances of survival (Veerappan et al., 2023). The presentation of evidence demonstrating the significant predictive potential of a biomarker is required to take place in prospective randomized clinical research. On the other hand, a predictive biomarker is able to provide information regarding the future efficacy of a treatment (Saravanan et al., 2023). Therefore, a predictive biomarker can assist in screening for a subgroup of patients who will respond favorably to a particular treatment (Merouane & Said, 2022). Because a predictive biomarker offers varied advantages depending on sub-patient risk groups indicated by the biomarker status (Khan & Shedole, 2022). This is because a biomarker can be used to predict the outcomes of clinical trials (Nave & Elbaz, 2021).

Proposed Work

The proposed method involves the development of prognostic models for the survivability of breast cancer. This method enhances the precision of prediction using a prognostic indicators and this include biochemical data and clinicopathological characteristics as in Figure 1.

Figure 1: Proposed Framework

Data Collection and Preprocessing:

In order for machine learning modeling to take place, the dataset must first go through the process of data preparation

Proposed DT algorithm

Data mining algorithms are utilized in the construction of decision tree-based prognostic models. Each algorithm is taught on the dataset in order to make their predictions regarding the path and prognosis of BC. Using ML, the hyperparameters of the machine learning models are fine-tuned.

Decision Support System (DSS):

ML-proposed models that have performed very well are incorporated into a DSS. The Decision Support System (DSS) uses a voting mechanism that integrates predictions from a number of different models in order to increase the overall accuracy and robustness of forecasts.

Data Collection and Preprocessing

Prior developing prognostic models, it is essential in collecting dataset consisting of breast cancer patients as in Table 1.

Table 1: Parameters of prognostic model

Parameter Number	Parameter Name	Description
1	Age	Age of the patient at diagnosis
2	Gender	Typically female for breast cancer
3	Tumor Size	Tumor size in centimeters
4	Lymph Node Involvement	Number of lymph nodes involved
5	Histological Grade	Grade of the tumor based on microscopic examination
6	Metastasis Presence	Presence or absence of metastasis (0 for no, 1 for yes)
7	Estrogen Receptor (ER) Status	Status of estrogen receptors (0 for negative, 1 for positive)
8	Progesterone Receptor (PR) Status	Status of progesterone receptors (0 for negative, 1 for positive)
9	HER2 Expression	Status of HER2 protein overexpression (0 for negative, 1 for positive)
10	Proliferation Index	Percentage of tumor cells positive for Ki67, indicating cell proliferation rate
11	Insulin Level	Insulin Level in the blood
12	Glucose Level	Glucose Level in the blood
13	Metabolite Levels	Levels of various related metabolites
14	Missing Value Imputation Method	Method used for handling missing values (e.g., mode imputation)
15	Normalization Method	Method used for normalizing continuous variables (e.g., min-max normalization)
16	Scaling Method	Method used for scaling features (e.g., standardization to mean 0, standard deviation 1)
17	Encoding Method for Categorical	Method used for encoding categorical variables
18	Correlation Threshold	Threshold for correlation analysis to select or eliminate features
19	Feature Importance Ranking Method	Method for ranking the importance of features (e.g., based on information gain)
20	Decision Tree Max Depth	Stopping criteria

1. **Clinicopathological Features:**
    - **Demographic Information**: The patient's age, gender are included in the personal details.
    - **Tumor Characteristics**: Presence or absence of metastases.
    - **Hormone Receptor Status**: It helps in determining the treatment plan that will be implemented.
    - **HER2/Neu Expression**: An overexpressed HER2 protein is associated to a more aggressive subtype of breast cancer and is responsible for dictating the treatment options that are available.
    - **Ki67 Proliferation Index**: This marker is helpful for identifying how aggressive a tumor is since it indicates the rate at which cancer cells are growing.
2. **Biochemical Data:**
    - **Glucose Metabolism Markers**: insulin, glucose, and related metabolite levels are becoming increasingly implicated in the development of cancer. These markers of glucose metabolism are also known as glucose metabolism markers.

Preprocessing:

The data is gathered and is used for training the machine learning models via a series of preprocessing stages.

1. **Handling Missing Values:**
    - A statistical procedure is used to fill in missing data is imputation using mode imputation.
2. **Normalization and Scaling:**
    - **Normalization**: Continuous variables (such glucose levels or tumor size) are normalized, which is often between 0 and 1 to guarantee.
    - **Scaling**: The features are modified using Support Vector Machine.

3. **Encoding Categorical Variables:**
    - **One-Hot Encoding**: This converts variables into vectors that machine learning systems are able to handle more effectively.
    - **Label Encoding**: An method for assigning a numerical value to each category is label encoding.
4. **Feature Selection:**
    - **Correlation Analysis**: Therefore, it is required to do a correlation analysis in order to detect and eliminate characteristics.
    - **Importance Ranking**: The process of ranking features with relevance to the prediction task is referred to as importance ranking.

Proposed Decision Tree Algorithm

Start with the entire dataset D.

$D=\{(x1,y1),(x2,y2),\ldots,(xn,yn)\}$

where

xi - feature vector and

yi - corresponding target variable.

2. Calculate Initial Entropy

Compute the dataset entropy to measure the disorder or impurity.

$H(D)=-\sum_{c=1}^{C} p(c)\log_2 p(c)$

where p(c) is the proportion of instances in class c, and C is the total number of classes.

3. Split the Dataset

Step: Perform the split on the feature A and value v that provides the highest information gain.

$D=Dv \cup D'v$

where Dv and D'v are the subsets of D where feature A takes the value v and does not take the value v, respectively.

4. Recursively Build the Tree

Step: Repeat steps 2-4 for each subset Dv and D'v until one of the stopping criteria.

Stopping Criteria:

- All instances in the subset belong to the same class.
- No remaining features to split on.
- Maximum depth of the tree is reached.

5. Assign Class Labels

```
function BuildDecisionTree(D, depth):
   if stopping criteria met:
      return leaf node with majority class label

   A_best, v_best = FindBestSplit(D)
   D_v, D_neg_v = SplitDataset(D, A_best, v_best)

   left_subtree
   right_subtree

   return node with A_best, v_best, left_subtree, right_subtree

function FindBestSplit(D):
   max_info_gain = -infinity
   best_feature = None
   best_value = None

   for each feature A in D
      for each value v in values(A)
function InformationGain(D, D_v, D_neg_v):
   return Entropy(D) - (|D_v| / |D| * Entropy(D_v) + |D_neg_v| / |D| * Entropy(D_neg_v))

function Entropy(D):
   H = 0
   for each class c in D:
      p = proportion of instances in class c
```

```
        H -= p * log2(p)
    return H

function SplitDataset(D, feature, value):
    D_v = subset of D where feature == value
    D_neg_v = subset of D where feature != value
    return D_v, D_neg_v
```

Decision Support System (DSS)

To construct a DSS is to combine the predictions of multiple machine learning models.

$y'i=f1(xi)\oplus f2(xi)\oplus\ldots\oplus fn(xi)$

where:

$y'i$ - combined prediction

$fj(xi)$ - ML model prediction

$\oplus$ - combining operation, which is voting.

2. Weighted Ensemble Method

Step 2: In order to create a weighted ensemble is to assign a weight to each predictor based on how well it performs.

$y'i=\sum_{j=1} w_j f_j(xi)$

where:

wj - weight assigned to the j^{th} predictor, satisfying $\sum_{j=1} w_j=1$.

3. Decision Rule Based on Risk Scores

Step 3: In order to determine a patient's risk score, the first step is to combine all of the forecasts with the cutoffs that have been initially defined.

Risk Score$i=\sum_{j=1}$ wjfj(xi)

{1 if Risk Score $i\geq\tau$

{0 if Risk Score $i\geq\tau$

where:

τ - threshold for classifying patients into high-risk (1) or low-risk (0) groups.

RESULTS AND DISCUSSION

Generally, the dataset is split, with a ratio of either 80/20 or 70/30. Following the training and fine-tuning, the testing set is used to test the models on new data that is unknown to the model. Through the utilization of cross-validation strategies such as 10-fold cross-validation, it is possible to get a model performance that is both resilient and generalizable across different data subsets.

Table 1. Results of Accuracy of various Machine Learning in training set.

Datasets	AUC	Sensitivity	Specificity	+LR	−LR
10	0.766	66.089	87.068	5.713	0.364
20	0.757	64.808	86.674	5.407	0.384
30	0.755	66.089	85.098	4.846	0.374
40	0.748	64.808	84.704	4.609	0.394
50	0.748	64.808	84.704	4.609	0.394
60	0.744	64.808	83.818	4.353	0.394
70	0.742	64.808	83.424	4.235	0.394
80	0.737	63.528	83.818	4.265	0.414
90	0.728	60.869	84.704	4.334	0.433
100	0.711	58.308	83.818	3.920	0.473

The majority of the predictors was more than or equal to 0.7, which is typically considered to be a threshold that is clinically important. From this group, 100 datasets was selected because it ranked highest for characteristics linked to glucose metabolism (Table 2), which are believed to have a substantial role in the advancement of breast cancer. The degree to which each group of training set attributes is more important than the others is given in Table 2.

Table 2. Attribute Groups Weights

Method		Averaging Weights					Normalized Weights				
		SVM	ANN	DNN	RNN	Proposed	SVM	ANN	DNN	RNN	Proposed

10	Training	0.413	1.030	0.594	0.334	0.581	0.138	0.344	0.198	0.111	0.194
20		0.761	1.833	1.373	0.891	0.992	0.128	0.309	0.231	0.150	0.167
30		0.421	0.900	1.148	0.387	0.579	0.121	0.258	0.329	0.111	0.166
40		0.442	1.263	0.621	0.437	0.526	0.132	0.378	0.186	0.131	0.158
10	Test set	0.455	1.160	0.549	0.336	0.469	0.151	0.385	0.182	0.112	0.156
20		0.539	1.379	0.781	0.582	0.601	0.137	0.350	0.198	0.148	0.152
30		0.633	1.115	0.355	0.389	0.446	0.212	0.374	0.119	0.130	0.149

By combining the two predictors in a DSS model for the advancement of breast cancer, whether it was both positive and negative, a c-statistic of 0.84 is formed with a 95% confidence interval as in Table 3.

Table 3. ML Performance

Parameter	Training	Test	Validation
F-measure	0.696	0.677	0.698
Accuracy	0.853	0.838	0.860
AUC	0.822	0.813	0.815
(+) LR	9.1	8.5	8.6
(−)LR	0.4	0.4	0.4
HR	10.7	10.3	10.9

LR is for "likelihood ratio," C.I. stands for "confidence interval," and HR stands for "hazard ratio." The following concepts are defined: a. Following the completion of a risk assessment that utilized both predictors, the analytical performance was evaluated after that.

Both the Cox proportional analysis and the ROC curve or not the combined DSS was able to differentiate between patients who experienced recurrence and those who did not, the AUC was computed on a risk scale that consisted of three levels: two, one, and zero.

Both the Kaplan-Meier and log-rank approaches were utilized in order to compute the survival curves through the utilization of software tools. From the time of recruitment until the progression of the disease, the progression-free survival (PFS) was calculated, which served as the endpoint of the experiment. The patients who did not exhibit any indicators of illness development were excluded from the study during the most recent follow-up.

In order to ascertain the probability of the development of breast cancer, Bayesian analysis was conducted along with the utilization of likelihood ratios, which included both positive (LR) and negative (−LR) values.

Inferences

Using a support vector machine (SVM) to develop an AI-based DSS for the purpose of prognostic evaluation of non-metastatic breast cancer patients has been successful, as illustrated here. To be more specific, a set of prognostic discriminators consisting of biochemical data and commonly gathered clinicopathological characteristics of BC patients might be constructed using the integration of machine learning and robotic otolaryngology techniques.

This combination strategy has the potential to increase the accuracy by assigning different weights to the various attributes associated with the model. Additionally, integrating ML and reinforcement learning methodologies results in a model that is easier to interpret and can be trained with smaller datasets. This is analogous to the way Bayesian networks were utilized for Bayesian classification. To add hitherto unanalyzed prognostic and metabolic parameters to the newly constructed DSS, which could all be easily extracted by EHR, machine learning has the potential to provide personalized therapy with large and long-lasting benefits. These advantages are expected to accrue to personalized medicine. It's possible that this will occur without driving up the expense of healthcare.

Limitations

A single establishment was the only one that participated in the research. The second problem is that the large sample size may have prevented machine learning from being as effective as it may have been. In spite of this, we believe that high-dimensional electronic health record data, when combined with machine learning algorithms and proposed models, has the potential to deliver predictive information and completely revolutionize personalized therapy.

CONCLUSION

In the event that we combine machine learning methods with proposed models, it is feasible that we will be able to better weight the relative relevance of attributes, which would ultimately result in increased model precision. In accordance with the prevalent pattern, the model that we have proposed strives to achieve both decision-making and model interpretability. This, in conjunction with the fact that we have utilized a real-world BC dataset, is the innovative aspect of our research. In order for any machine learning solution to be put into clinical practice, it must first be validated through prospective studies that involve several centers and adequately address any privacy concerns that are associated with digital electronic health record data.

REFERENCES

[1] Tran, W. T., Jerzak, K., Lu, F. I., Klein, J., Tabbarah, S., Lagree, A., & Sadeghi-Naini, A. (2019). Personalized breast cancer treatments using artificial intelligence in radiomics and pathomics. Journal of Medical Imaging and Radiation Sciences, 50(4), S32-S41. https://doi.org/10.1016/j.jmir.2019.05.001

[2] Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. Breast Cancer Research, 21, 1-11. https://doi.org/10.1186/s13058-019-1178-0

[3] Ahn, J. S., Shin, S., Yang, S. A., Park, E. K., Kim, K. H., Cho, S. I., & Kim, S. (2023). Artificial intelligence in breast cancer diagnosis and personalized medicine. Journal of Breast Cancer, 26(5), 405. https://doi.org/10.4048/jbc.2023.26.e38

[4] Vadapalli, S., Abdelhalim, H., Zeeshan, S., & Ahmed, Z. (2022). Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Briefings in Bioinformatics, 23(5), bbac191. https://doi.org/10.1093/bib/bbac191

[5] Low, S. K., Zembutsu, H., & Nakamura, Y. (2018). Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Science, 109(3), 497-506. https://doi.org/10.1111/cas.13474

[6] Rezayi, S., Niakan Kalhori, S. R., & Saeedi, S. (2022). Effectiveness of artificial intelligence for personalized medicine in neoplasms: A systematic review. BioMed Research International, 2022, Article 9497461. https://doi.org/10.1155/2022/9497461

[7] Alzu'bi, A., Najadat, H., Doulat, W., Al-Shari, O., & Zhou, L. (2021). Predicting the recurrence of breast cancer using machine learning algorithms. Multimedia Tools and Applications, 80(9), 13787-13800. https://doi.org/10.1007/s11042-021-10207-y

[8] Sugimoto, M., Hikichi, S., Takada, M., & Toi, M. (2023). Machine learning techniques for breast cancer diagnosis and treatment: A narrative review. Annals of Breast Surgery, 7. https://doi.org/10.21037/abs-23-07

[9] Manikis, C., Simos, N. J., Kourou, K., Kondylakis, H., Poikonen-Saksela, P., Mazzocco, K., & Fotiadis, D. (2023). Personalized risk analysis to improve the psychological resilience of women undergoing treatment for breast cancer: Development of a machine learning–driven clinical decision support tool. Journal of Medical Internet Research, 25, e43838. https://doi.org/10.2196/43838

[10] Sotudian, S., & Paschalidis, I. C. (2021). Machine learning for pharmacogenomics and personalized medicine: A ranking model for drug sensitivity prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(4), 2324-2333. https://doi.org/10.1109/TCBB.2021.3072687

[11] Saravanan, V., Parameshachari, B. D., Hussein, A. H. A., Shilpa, N., & Adnan, M. M. (2023, November). Deep learning techniques based secured biometric authentication and classification using ECG signal. In 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS) (pp. 1-5). IEEE. https://doi.org/10.1109/ICIICS56722.2023.10206927

[12] Yuvaraj, N., Praghash, K., Arshath Raja, R., Chidambaram, S., & Shreecharan, D. (2022, December). Hyperspectral image classification using denoised stacked auto encoder-based restricted Boltzmann machine classifier. In International Conference on Hybrid Intelligent Systems (pp. 213-221). Springer. https://doi.org/10.1007/978-3-031-16589-1_18

[13] Veerappan, K. N. G., Natarajan, Y., Raja, A., Perumal, J., & Kumar, S. J. N. (2023). Categorical data clustering using meta heuristic link-based ensemble method: Data clustering using soft computing techniques. In Dynamics of Swarm Intelligence Health Analysis for the Next Generation (pp. 226-238). IGI Global. https://doi.org/10.4018/978-1-7998-7641-2.ch014

[14] Saravanan, V., Sankaradass, V., Shanmathi, M., Bhimavarapu, J. P., Deivakani, M., & Ramasamy, S. (2023, May). An early detection of ovarian cancer and the accurate spreading range in the human body by using deep medical learning model. In 2023 International Conference on Disruptive Technologies (ICDT) (pp. 68-72). IEEE. https://doi.org/10.1109/ICDT56184.2023.10001382

[15] Merouane, E., & Said, A. (2022). Prediction of metastatic relapse in breast cancer using machine learning classifiers. International Journal of Advanced Computer Science and Applications, 13(2). https://doi.org/10.14569/IJACSA.2022.0130241
[16] Khan, D., & Shedole, S. (2022). Leveraging deep learning techniques and integrated omics data for tailored treatment of breast cancer. Journal of Personalized Medicine, 12(5), 674. https://doi.org/10.3390/jpm12050674
[17] Nave, O., & Elbaz, M. (2021). Artificial immune system features added to breast cancer clinical data for machine learning (ML) applications. Biosystems, 202, Article 104341. https://doi.org/10.1016/j.biosystems.2021.104341