Ensemble Machine-Learning Models for Breast Cancer Identification

Elias Dritsas[1], Maria Trigka[2], and Phivos Mylonas[2]

[1] Department of Electrical and Computer Engineering, University of Patras, Greece
dritsase@ceid.upatras.gr
[2] Department of Informatics and Computer Engineering, University of West Attica, Greece
{mtrigka,mylonasf}@uniwa.gr

Abstract. The advances in the Machine Learning (ML) domain, from pattern recognition to computational statistical learning, have increased its utility for breast cancer as well by contributing to the screening strategy of diverse risk factors with complex relationships and personalized early prediction. In this work, we focused on Ensemble ML models after using the synthetic minority oversampling technique (SMOTE) with 10-fold cross-validation. Models were compared in terms of precision, accuracy, recall and area under the curve (AUC). After the experimental evaluation, the model that prevailed over the others was the Rotation Forest (RotF) achieving accuracy, precision and recall equal to 82% and an AUC of 87.4%.

Keywords: Ensemble Models · Machine-Learning · Risk-Prediction.

1 Introduction

Breast cancer develops from cells/tissue in the breast gland. It is the most common type of cancer that occurs in women in both developed and less developed countries. It is the most common malignancy in women with an incidence of 12%, i.e. 1 in 8, and the second most common cause of cancer death after lung cancer [1, 20].

Although breast cancer also occurs in men, it is very rare. The incidence of male breast cancer is less than 1% of all breast cancer cases. In 2020, there were 2.3 million women diagnosed with breast cancer and 685000 deaths globally, according to the World Health Organization (WHO) [1, 23].

Every woman needs to be aware of the aggravating factors that increase the risk of developing the disease, the value of prevention, self-examination, and early diagnosis, as well as the effectiveness of modern treatment. Age is a factor in the occurrence of breast cancer, as most cases occur after the age of 50, while it is rare in women under the age of 35. In addition, women who have already been diagnosed with cancer are more likely to develop new cancer in the same or the other breast. Also, women in whom menstruation began at the age of

fewer than 12 years or stopped at the age of more than 55 years have a relatively increased risk of developing the disease [26, 34, 8, 37].

Obese women have a higher risk of being diagnosed with breast cancer compared to women who maintain a healthy weight. Alcohol consumption and smoking are also aggravating factors as are estrogen and progesterone. Finally, research has shown that women with dense breasts have an increased chance of developing cancer [28, 29, 22].

In the initial stage, breast cancer shows no symptoms. A palpable mass, change in skin colour, infiltration or discharge may later appear. If a woman does not pay attention to the aforementioned symptoms, then she may show signs of advanced disease, such as redbreast (inflammatory cancer), bone pain, and large swelling. The diagnosis of breast cancer in the first stage or even in a pre-cancerous stage is much more due to the awareness of women regarding the preventive control of the breasts with clinical palpation by a doctor, mammography and ultrasound, as well as with self-palpation. Once a suspicious tumour is found, the diagnosis is made by taking material from the tumour for microscopic examination [31, 41].

In the second half of the 20th century and up to today, there have been rapid developments in the knowledge and treatment of breast cancer. The introduction of screening healthy women with mammography dramatically changes the profile of the disease and its outcome. New technologies are being added to breast cancer imaging and diagnosis. The discovery of more and more biomarkers decodes the heterogeneity of breast cancer, which is now classified into different groups with different prognostic models and methods. Finally, October 25 is a day that concerns all women, since it is dedicated to their fight against breast cancer [25, 20, 4].

Machine learning has played an important role in the medical field as it contributes to the early prediction of various diseases complications, such as diabetes (as classification [14] or regression task for continuous glucose prediction [11]), cholesterol [21], hypertension [12], chronic obstructive pulmonary disease [10], covid-19 [18], stroke [17], chronic kidney disease [16], cardiovascular diseases [19], lung cancer [15], and metabolic syndrome [13] etc.

In this research work, we relied on anthropometric data and biochemical indices, which can be collected in routine blood analyses to predict the occurrence of breast cancer. The main contribution of this study lies in the selected models for evaluation. Ensemble machine learning models were assessed based on several predictors that can potentially be used as breast cancer biomarkers. Moreover, from a methodological perspective, before the models' evaluation, we employed SMOTE to render the dataset balanced which, as will be verified in the experiments, favoured the classifiers' performance. Finally, comparing our outcomes with the ones derived from a previous study in the same dataset, it is shown that ensemble methods constitute an alternative and highly efficient solution for breast cancer prediction.

The remainder of the paper is organized as follows. In Section 2, a description of the adopted methodology is outlined. Furthermore, in Section 3, we discuss

related works on the topic under consideration and note the research results. Finally, conclusions and future directions are presented in Section 4.

2 Methodology

In this section, the dataset and its characteristics are described, the adopted methodology is noted, the models have been described as well as the evaluation metrics with which the experimental evaluation was carried out.

2.1 Dataset Presentation and Processing

A detailed analysis of the dataset, the adopted methodology for the measurements capturing and the determination of the class label per subject have been made by the authors in [35]. The dataset we relied on comes from the UCI Machine Learning Repository [2]. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. Specifically, the attributes are Age (years), Body Mass Index (BMI) (kg/m^2), Glucose (mg/dL), Insulin (μU/mL), Homeostasis Model Assessment (HOMA), Leptin (ng/mL), Adiponectin (μg/mL), Resistin (ng/mL) and Monocyte Chemoattractant Protein-1 (MCP-1) (pg/dL). The total number of participants is 116, of which 64 (or 55.2%) have been diagnosed with breast cancer and 52 (or 44.8%) have not. Statistical details about the features in the given dataset are illustrated in Table 1.

Attribute	**Description**		
	Min	**Max**	**Mean±stdDev**
Age	24	89	57.3±16.1
BMI	18.37	38.58	27.58±5.02
Glucose	60	201	97.79±22.52
Insulin	2.432	58.46	10.01±10.07
HOMA	0.467	25.05	2.69±3.64
Leptin	4.311	90.28	26.61±19.18
Adiponectin	1.656	38.04	10.18±6.84
Resistin	3.21	82.1	14.73±12.39
MCP-1	45.843	1698.44	534.65±345.91

Table 1. Statistical description of the features in the dataset.

Here, for data analysis and knowledge extraction, we experimented with a free software tool, the Waikato environment (Weka), widely exploited for data preprocessing and predictive modelling [3]. Concerning data preprocessing, there were no missing values so any data imputation technique wasn't applied. Moreover, investigating the considered features-predictors no outliers were detected. Then, we applied SMOTE [9] technique to make the class distribution uniform.

The motivation behind the application of SMOTE is to further improve the classifiers' discrimination ability and sensitivity in identifying both patients and healthy subjects. SMOTE builds iteratively a new class-balanced dataset exploiting a portion of the minority class data to make the class distribution 50-50%. SMOTE is preferable since it doesn't create duplicates but new synthetic data using the k-NN method. Given that we worked on Weka, we have used the offered method where $k = 5$. An algorithmic overview of SMOTE is shown in Algorithm 1. Note that the statistical description of the balanced dataset was omitted to be referred to since there was no significant difference from the information presented in Table 1.

Algorithm 1 SMOTE

Input: M (minority class sample size), N (% of synthetic minority samples for class balancing), $k = 5$ (number of nearest neighbors), s_{syn} synthetic instance;
Choose randomly a subset $\mathcal{S}$ of the minority class data of size $S = \frac{N}{100}M$ (synthetic minority data ratio) such that the class labels are balanced;
for all $s_i \in \mathcal{S}$ **do**
(1) Find the $k = 5$ nearest neighbors;
(2) Randomly select one of the $k = 5$ NNs, called $\hat{s}_i$;
(3) Calculate the distance $d_{i,k} = \hat{s}_i - s_i$ between the randomly selected NN $\hat{s}_i$ and the instance s_i;
(4) The new synthetic instance is generated as $s_{syn} = s_i + \delta d_{i,k}$ (where $\delta = rand(0, 1)$ is a random number between 0 and 1);
end for
Repeat steps number 2–4 until the desired proportion of minority class is met.

2.2 Machine Learning Models and Evaluation Metrics

In this research work, we focused on ensemble models [39], which is a machine learning approach that combines multiple other models in the prediction process. More specifically, the Bagging [27] model was configured to have as a base classifier the Random Forest (RF) [30], the Stacking [40] method was set to combine the base classifiers RF and J48 [36] model, and as a meta classifier the Logistic Regression (LR) [33] model and the Voting [27] method considered the same based models with the stacking, but at the final step averages the probabilities to predict the class (soft voting). Finally, Rotation Forest [38] was set up to use Principal Component Analysis for features transformation and the RF as a base classifier.

To evaluate the performance of ML models, we applied 10-fold cross-validation and relied on metrics commonly used in the ML field, namely accuracy, precision, recall, and AUC. The definition of these metrics is based on the confusion matrix consisting of the elements true positive(Tp), true negative (Tn), false positive (Fp) and false-negative (Fn). Hence, the aforementioned metrics are defined as follows:

– Accuracy:

$$\text{Accuracy} = \frac{\text{Tn} + \text{Tp}}{\text{Tn} + +\text{Fn} + \text{Tp} + \text{Fp}} \tag{1}$$

– Precision:

$$\text{Precision} = \alpha_1 \frac{\text{Tp}}{\text{Tp} + \text{Fp}} + \alpha_2 \frac{\text{Tn}}{\text{Tn} + \text{Fn}} \tag{2}$$

– Recall:

$$\text{Recall} = \alpha_1 \frac{\text{Tp}}{\text{Tp} + \text{Fn}} + \alpha_2 \frac{\text{Tn}}{\text{Tn} + \text{Fp}} \tag{3}$$

– To evaluate the distinguishability of a model, the AUC is exploited. It is a metric that varies in [0, 1].

It should be noted that (2), (3) capture the weighted average precision and recall. In both equations, the first term concerns the class "Yes" while the latter relates to the class "No". Generally, if the dataset instances are distributed non-uniformly among the two classes the weights $\alpha_1 = 44.8\%, \alpha_2 = 55.2\%$ will be different and specifically $\alpha_1 \neq \alpha_2 \neq 50\%$. In this study, we have used SMOTE technique to acquire a balanced dataset, therefore, $\alpha_1 = \alpha_2 = 50\%$.

3 Results and Discussion

This section provides a brief overview of the related works on the topic under consideration and notes our experimental results. Observing Table 2, the most common ensemble models are assessed in terms of accuracy, precision, recall and AUC. Also, in the context of our analysis, the selected models were evaluated before and after the application of class balancing using SMOTE. As the results witnessed, the use of SMOTE raised the models' performance metrics by around 5%. The suggested model is Rotation Forest (after SMOTE) which indicated Accuracy, Precision and Recall of 82% and AUC of 87.4%.

Ensemble Models	**Accuracy**		**Precision**		**Recall**		**AUC**	
	Balanced	**UnBalanced**	**Balanced**	**UnBalanced**	**Balanced**	**UnBalanced**	**Balanced**	**UnBalanced**
Random Forest	0.773	0.724	0.774	0.723	0.773	0.724	0.867	0.807
Voting	0.805	0.698	0.805	0.697	0.805	0.698	0.867	0.800
Bagging	0.804	0.767	0.805	0.768	0.805	0.767	0.870	0.822
Stacking	0.818	0.741	0.816	0.741	0.816	0.741	0.872	0.806
Rotation Forest	0.820	0.775	0.820	0.775	0.820	0.776	0.874	0.824

Table 2. Performance Evaluation of Ensemble Models.

Now, let us focus on a recent study that experimented with the same dataset. The proposed model in study [6] is Support Vector Machine (SVM) combined with an extra-trees model for feature selection that performed Accuracy equal to 80.23%, Precision and Recall of 82.71% and 78.57%, correspondingly, and

an AUC of 78%. Comparing the prevailing model of work [6] with the suggested model in this study, it is shown the prevalence of ensemble model Rotation Forest against SVM in all metrics.

At this point, we will pay attention to works that study breast cancer exploiting different datasets from the ones considered above. In the paper [7], the authors compare the predictive accuracy of the Naive Bayes (NB) classifier and K-Nearest Neighbour (KNN) for breast cancer classification using cross-validation. The experimental results showed that the KNN gives the highest accuracy (97.51%). Moreover, in [5], the authors implemented three machine learning models, namely Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN), for predicting breast cancer recurrence and comparing them in terms of accuracy, sensitivity and specificity. The SVM model prevailed in all the aforementioned metrics.

Similarly, in [24], the authors compared five supervised ML models, namely SVM, KNN, RF, ANN and LR. The results revealed that the ANN achieved the highest accuracy, precision, and F1 score of 98.57%, 97.82%, and 0.98, respectively, whereas 97.14%, 95.65%, and 0.97 accuracy, precision, and F1 score are obtained by the SVM, respectively. Finally, in [32], a performance comparison is performed between SVM, DT, NB, and KNN models on the Wisconsin Breast Cancer dataset for breast cancer risk prediction using the WEKA tool. The experimental results showed that SVM achieved the highest accuracy of 97.13% with the lowest error rate.

4 Conclusions

In the context of this work, we focused on Ensemble ML models, namely RF, Stacking, Bagging, Voting and RotF, to accurately capture the probability of breast cancer occurrence based on critical biochemical indexes. Our models were compared based on the accuracy, precision, recall and AUC metrics to reveal the most suitable one for distinguishing between patients and non-patients. The experimental results showed that the RotF model prevailed over the others achieving accuracy, precision and recall equal to 82% and an AUC of 87.4% after the SMOTE technique with 10-fold cross-validation.

In future work, we intend to follow an alternative path for detecting cancerous tumours by focusing on X-ray images and, thus exploiting efficient processing techniques from Computer Vision, Image Processing and Deep Learning.

References

1. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer ((accessed on 1 April 2023))
2. Uci machine learning repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra ((accessed on 1 April 2023))
3. Weka. https://www.weka.io/ ((accessed on 1 April 2023))

4. Ahmad, A.: Breast cancer statistics: recent trends. Breast cancer metastasis and drug resistance: challenges and progress pp. 1–7 (2019)
5. Ahmad, L.G., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., Razavi, A., et al.: Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform **4**(124), 3 (2013)
6. Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N.L., Atmaji, F.T.D., Widodo, T., Bahiyah, N., Benes, F., Rhee, J.: Predicting breast cancer from risk factors using svm and extra-trees-based feature selection method. Computers **11**(9), 136 (2022)
7. Amrane, M., Oukid, S., Gagaoua, I., Ensari, T.: Breast cancer classification using machine learning. In: 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT). pp. 1–4. IEEE (2018)
8. Billena, C., Wilgucki, M., Flynn, J., Modlin, L., Tadros, A., Razavi, P., Braunstein, L.Z., Gillespie, E., Cahlon, O., McCormick, B., et al.: 10-year breast cancer outcomes in women $\leq$ 35 years of age. International Journal of Radiation Oncology* Biology* Physics **109**(4), 1007–1018 (2021)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
10. Dritsas, E., Alexiou, S., Moustakas, K.: Copd severity prediction in elderly with ml techniques. In: Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 185–189 (2022)
11. Dritsas, E., Alexiou, S., Konstantoulas, I., Moustakas, K.: Short-term glucose prediction based on oral glucose tolerance test values. In: HEALTHINF. pp. 249–255 (2022)
12. Dritsas, E., Alexiou, S., Moustakas, K.: Efficient data-driven machine learning models for hypertension risk prediction. In: 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). pp. 1–6. IEEE (2022)
13. Dritsas, E., Alexiou, S., Moustakas, K.: Metabolic syndrome risk forecasting on elderly with ml techniques. In: Learning and Intelligent Optimization: 16th International Conference, LION 16, Milos Island, Greece, June 5–10, 2022, Revised Selected Papers. pp. 460–466. Springer (2023)
14. Dritsas, E., Trigka, M.: Data-driven machine-learning methods for diabetes risk prediction. Sensors **22**(14), 5304 (2022)
15. Dritsas, E., Trigka, M.: Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing **6**(4), 139 (2022)
16. Dritsas, E., Trigka, M.: Machine learning techniques for chronic kidney disease risk prediction. Big Data and Cognitive Computing **6**(3), 98 (2022)
17. Dritsas, E., Trigka, M.: Stroke risk prediction with machine learning techniques. Sensors **22**(13), 4670 (2022)
18. Dritsas, E., Trigka, M.: Supervised machine learning models to identify early-stage symptoms of sars-cov-2. Sensors **23**(1), 40 (2022)
19. Dritsas, E., Trigka, M.: Efficient data-driven machine learning models for cardiovascular diseases risk prediction. Sensors **23**(3), 1161 (2023)
20. Fahad Ullah, M.: Breast cancer: current perspectives on the disease status. Breast Cancer Metastasis and Drug Resistance: Challenges and Progress pp. 51–64 (2019)
21. Fazakis, N., Dritsas, E., Kocsis, O., Fakotakis, N., Moustakas, K.: Long-term cholesterol risk prediction using machine learning techniques in elsa database. In: IJCCI. pp. 445–450 (2021)

22. Gordon, P.B.: The impact of dense breasts on the stage of breast cancer at diagnosis: A review and options for supplemental screening. Current Oncology **29**(5), 3595–3636 (2022)
23. Gucalp, A., Traina, T.A., Eisner, J.R., Parker, J.S., Selitsky, S.R., Park, B.H., Elias, A.D., Baskin-Bey, E.S., Cardoso, F.: Male breast cancer: a disease distinct from female breast cancer. Breast cancer research and treatment **173**, 37–48 (2019)
24. Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M., Kabir, M.N.: Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science **1**, 1–14 (2020)
25. Jafari, S.H., Saadatpour, Z., Salmaninejad, A., Momeni, F., Mokhtari, M., Nahand, J.S., Rahmati, M., Mirzaei, H., Kianmehr, M.: Breast cancer diagnosis: Imaging techniques and biochemical markers. Journal of cellular physiology **233**(7), 5200–5213 (2018)
26. Johansson, A.L., Trewin, C.B., Hjerkind, K.V., Ellingjord-Dale, M., Johannesen, T.B., Ursin, G.: Breast cancer-specific survival by clinical subtype after 7 years follow-up of young and elderly women in a nationwide cohort. International journal of cancer **144**(6), 1251–1261 (2019)
27. Kabari, L.G., Onwuka, U.C.: Comparison of bagging and voting ensemble machine learning algorithm as a classifier. International Journals of Advanced Research in Computer Science and Software Engineering **9**(3), 19–23 (2019)
28. Lee, K., Kruper, L., Dieli-Conwright, C.M., Mortimer, J.E.: The impact of obesity on breast cancer diagnosis and treatment. Current oncology reports **21**, 1–6 (2019)
29. Li, H., Terry, M.B., Antoniou, A.C., Phillips, K.A., Kast, K., Mooij, T.M., Engel, C., Noguès, C., Stoppa-Lyonnet, D., Lasset, C., et al.: Alcohol consumption, cigarette smoking, and risk of breast cancer for brca1 and brca2 mutation carriers: results from the brca1 and brca2 cohort consortium. Cancer Epidemiology, Biomarkers & Prevention **29**(2), 368–378 (2020)
30. Liu, Y., Wang, Y., Zhang, J.: New machine learning algorithm: Random forest. In: Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3. pp. 246–252. Springer (2012)
31. Mokhatri-Hesari, P., Montazeri, A.: Health-related quality of life in breast cancer patients: review of reviews from 2008 to 2018. Health and quality of life outcomes **18**, 1–25 (2020)
32. Naji, M.A., El Filali, S., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A., Debauche, O.: Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science **191**, 487–492 (2021)
33. Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y., Cheng, C.Y.: Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology **122**, 56–69 (2020)
34. Olsson, H.L., Olsson, M.L.: The menstrual cycle and risk of breast cancer: a review. Frontiers in oncology **10**, 21 (2020)
35. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., Caramelo, F.: Using resistin, glucose, age and bmi to predict the presence of breast cancer. BMC cancer **18**(1), 1–8 (2018)
36. Posonia, A.M., Vigneshwari, S., Rani, D.J.: Machine learning based diabetes prediction using decision tree j48. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). pp. 498–502. IEEE (2020)
37. Riggio, A.I., Varley, K.E., Welm, A.L.: The lingering mysteries of metastatic recurrence in breast cancer. British journal of cancer **124**(1), 13–26 (2021)

38. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence **28**(10), 1619–1630 (2006)
39. Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1249 (2018)
40. Satapathy, S.K., Bhoi, A.K., Loganathan, D., Khandelwal, B., Barsocchi, P.: Machine learning with ensemble stacking model for automated sleep staging using dual-channel eeg signal. Biomedical Signal Processing and Control **69**, 102898 (2021)
41. Wang, L.: Early diagnosis of breast cancer. Sensors **17**(7), 1572 (2017)