**Introduction**

Breast cancer is the second most common women’s cancer in the World, right after cancers of the skin [1]. Furthermore, breast cancer has recently surpassed lung cancers as the leading cause of cancer death in women [2]. About 95% of newly diagnosed breast cancer patients are diagnosed in early, non-metastatic stage of disease [3].

After the introduction of adjuvant and neoadjuvant therapies in the treatment of early breast cancer, the probability of 5 year survival almost doubled in last 50 years [4]. Prognostic and predictive factors are crucial in the process of defining an optimal treatment strategy. One of the most important prognostic factors is the positivity of axillar lymph nodes [6]. Due to a recent change in the treatment strategy, significantly more patients are treated with neoadjuvant systemic therapy (NST), based on the tumour biology characteristics and radiological findings [7]. It is well known that axillary lymph node status is not precisely determined by radiological tests, ending up with close to 30% patients misdiagnosed (ref). Consequently, after neoadjuvant therapy, the status of sentinel lymph node biopsy (SLNB) as the most common primary method used to exclude axillary metastases is sometimes misinterpreted due to metastases responding to NST [8]. The neoadjuvant concept allows in vivo testing of treatment sensitivity, further personalization of the adjuvant part of systemic therapy, and provides the way to receive an accelerated approval of new therapies. It is also a valid model for the development of predictive biomarkers and reduces patient numbers in clinical trials by the usage of new surrogate endpoints such as pathological complete response (pCR) (ref).

Despite all the breakthroughs in our understanding of breast tumor biology adjuvant/neoadjuvant therapy in early breast cancer therapy is still based on the stage of the disease, and lymph node positivity usually prevents de-escalation of systemic therapy or omittance of lymph node radiotherapy. Therefore, our knowledge about the tumour’s ability to metastasize and the true status of lymph nodes is of paramount importance in our decision-making process. As an example, the current standard of neoadjuvant therapy for HER2 positive disease is an nthracycline/taxane-based chemotherapy in combination with trastuzumab and pertuzumab. This is followed by breast surgery, adjuvant radiotherapy (if indicated), and completion of the HER2-directed therapy, and, depending on the tumor biology, endocrine adjuvant therapy. The type of adjuvant part of HER2-directed therapy, monotherapy with trastuzumab or dual therapy with trastuzumab and pertuzumab is based on the status of axillary lymph nodes positivity, where in the case of positive nodes dual therapy is given (Afinity study, NCCN).

Since NST is used more often, there is a need for supplementary methods to aid multidisciplinary teams in the assessment of metastatic lymph node status. Due to exponential growth in oncology patient data generation, “Data Science” and machine learning techniques are extensively researched and are applied as possible solutions to various clinical problems [9,10]. Machine learning, as a computational method that maps a mathematical function to a dataset of independent variables in order to predict/classify a dependent variable, differs from traditional programming in that it directly learns from the data, without the need for explicit step-wise programming [11]. Traditional machine learning algorithms, such as support vector machines (SVM) and random forests (RF), have been successfully used to classify breast cancer into triple negative and non-triple negative types, predict the metastatic status of patients, and aid in detecting early disease recurrence [7,8]. Furthermore, more complicated models such as the gradient boosted trees and eXtreme Gradient Boosting (XGBoost) were used in the prediction of survival outcomes in patients with epithelial ovarian cancer, and the prediction of metastatic status in breast cancer, respectively [12,13].

In this research, we developed and evaluated several machine learning models that were trained on multiple features obtained from demographic data, imaging studies and pathohistological data (patient age, tumour size, ki-67, ER, PR, HER-2, pathohistological type, and gradus), with a goal to predict patient lymph node status accurately.

**Methods**

**Data source and preparation**

Data examined in this study was collected from 25 Croatian hospitals that cover over 95 % of all Croatian breast cancer cases , during a five year period, from 01/2017 to 01/2022. The collected data consists of 10 features: patient age at diagnosis, tumour size (in cm), pathohistological type, immunophenotype, pathohistological gradus, estrogen (ER) and progesterone (PR) receptor quantities (0-100), HER-2 levels (0-3), ki-67 index (0-100), and lymph node metastasis status (0/1).

The case group was defined as patients with the evidence of breast cancer lymph node metastasis. Consequently, the control group was defined as patients without the evidence of lymph node metastasis. Tumour samples were obtained via surgery and needle biopsies, while the target variable ground truth (lymph node positivity) was established by post-surgical lymph node pathohistological examination.

Initial data contained 13581 entries, from which 3872 entries had various missing values, ranging from the target variable (lymph node metastasis status) to pathohistological type and gradus. After we omitted the missing values, we were left with 9709 entries with complete data.

Since the model’s target population are patients who would potentially receive neoadjuvant therapy, we identified those patients from our study population (all patients received surgical treatment) using the following criteria: 1) all tumours with size > 5 cm (irrespective to subtype), 2) tumours with size ≥2cm of triple-negative or HER-2 positive subtype, 3) tumours of inflammatory subtype . After applying the following criteria, the final number of patients was 796.

In addition to the model based only on patients who would potentially receive neoadjuvant treatment, we also developed a broader model that generalizes to our entire breast cancer population (n= 9709), to see if similar performances are obtained.

**Table 1.** Clinical features used for breast cancer lymph node metastasis prediction

|  |  |
| --- | --- |
| Clinical features |  |
| Demographic data | Age at diagnosis |
| Radiology data | Tumour size |
| Pathology data | Pathohistological type |
|  | Immunophenotype |
|  | Gradus |
|  | ER |
|  | PR |
|  | HER-2 |
|  | Ki-67 |

Abbreviations: ER- estrogen receptors, PR- progesterone receptors, HER-2- human epidermal growth factor receptor 2, Ki-67- marker of proliferation.

**Prediction model development**

Features presented in Table 1. were used in model development. We split the data into a training batch (90 % of data) and a test batch (15 % of data). Categorical variables were “dummy encoded”, which resulted in a final table with 24 features (columns). Further, we performed a stratified 5-fold cross validation on our training sample to develop our model. Next, we optimized the model's hyperparameters by combining stratified cross-validation and halving grid-search. In this manner, we evaluated Logistic Regression, Support Vector Machine Classifier, Random Forrest Classifier, and eXtreme Gradient Boosting (XGBoost) Classifier. Model development was performed on Python (version 3.9. 5, Python Software Foundation) using the following libraries: “pandas”, “scikit-learn”, and “xgboost”.

**Model evaluation**

Final evaluation and predictions were made on the test sample (10 % of data). ROC curve was plotted, and areas under the curve (AUC) were obtained for each model. The model with the highest AUC was selected for further investigation. The optimal cut-off points for sensitivity and specificity were based on the Youden Index [14].

**Feature importance analysis**

We assessed feature importance by determining the number of times a feature became a split feature, as well as by calculating the mean decrease in Gini impurity [15].

**Statistical analysis**

**….**

**Results**

**Patient characteristics**

In total, 6414 (66 %) patients were identified as controls, while 3295 (34 %) patients were identified as cases.

Table 2. Patient characteristics (case/control comparison)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lymph node metastasis  group | Non- lymph node metastasis group | P- value |
| Num. of patients | 3295 (34 %) | 6414 (66 %) |  |
| Age | 61.9 (13.5) | 61.8 (12.3) | 0.535 |
| Tumour size | 2.73 (2.04) | 1.71 (1.2) | < 0.001\* |
| Tumour gradus (mode) | 2 (59 %) | 2 (54 %) |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J Clin* **2020**, *70*, 7-30, doi:10.3322/caac.21590.

2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **2021**, *71*, 209-249, doi:<https://doi.org/10.3322/caac.21660>.

3. Centers for Disease Control and Prevention. Incidence and Relative Survival by Stage at Diagnosis for Common Cancers. USCS Data Brief, n.A., GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2021.

4. Nordenskjöld, A.E.; Fohlin, H.; Arnesson, L.G.; Einbeigi, Z.; Holmberg, E.; Albertsson, P.; Karlsson, P. Breast cancer survival trends in different stages and age groups – a population-based study 1989–2013. *Acta Oncologica* **2019**, *58*, 45-51, doi:10.1080/0284186X.2018.1532601.

5. Redig, A.J.; McAllister, S.S. Breast cancer as a systemic disease: a view of metastasis. *J Intern Med* **2013**, *274*, 113-126, doi:10.1111/joim.12084.

6. Kim, D.D.; Basu, A. Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. *Value Health* **2016**, *19*, 602-613, doi:10.1016/j.jval.2016.02.008.

7. Gradishar, W.; Salerno, K.E. NCCN Guidelines Update: Breast Cancer. *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw* **2016**, *14*, 641-644, doi:10.6004/jnccn.2016.0181.

8. Vila, J.; Mittendorf, E.A.; Farante, G.; Bassett, R.L.; Veronesi, P.; Galimberti, V.; Peradze, N.; Stauder, M.C.; Chavez-MacGregor, M.; Litton, J.F.; et al. Nomograms for Predicting Axillary Response to Neoadjuvant Chemotherapy in Clinically Node-Positive Patients with Breast Cancer. *Ann Surg Oncol* **2016**, *23*, 3501-3509, doi:10.1245/s10434-016-5277-1.

9. Kann, B.H.; Hosny, A.; Aerts, H. Artificial intelligence for clinical oncology. *Cancer Cell* **2021**, *39*, 916-927, doi:10.1016/j.ccell.2021.04.002.

10. Boehm, K.M.; Khosravi, P.; Vanguri, R.; Gao, J.; Shah, S.P. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* **2022**, *22*, 114-126, doi:10.1038/s41568-021-00408-3.

11. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science* **2021**, *2*, 160, doi:10.1007/s42979-021-00592-x.

12. Paik, E.S.; Lee, J.W.; Park, J.Y.; Kim, J.H.; Kim, M.; Kim, T.J.; Choi, C.H.; Kim, B.G.; Bae, D.S.; Seo, S.W. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. *J Gynecol Oncol* **2019**, *30*, e65, doi:10.3802/jgo.2019.30.e65.

13. Li, Q.; Yang, H.; Wang, P.; Liu, X.; Lv, K.; Ye, M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J Transl Med* **2022**, *20*, 177, doi:10.1186/s12967-022-03369-9.

14. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32-35, doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

15. Adams, K. *Tutorial The Gini Impurity index and what it means and how to calculate it*; 2018.