Predicting Quality of Life for Breast Cancer Patients*

Christos Raspoptsis, Eugenia Mylona, Konstantina Kourou, Georgios Manikis, Haridimos Kondylakis, Kostas Marias, Paula Poikonen-Saksela, Panagiotis Simos, Evangelos Karademas, Ketti Mazzocco, Ruth Pat-Horenczyk, Berta Sousa, Dimitrios I. Fotiadis

Abstract— **The diagnosis of breast cancer has a significant impact on a patient's quality of life. Several demographic and clinical factors have been reported to affect the quality of life of breast cancer patients. However, few studies have a sufficient sample size for multifactorial assays to be tested. In the present work, we explore a rich set of clinical, psychological, socio-demographic, and lifestyle data from a large multicenter study of breast cancer patients (n = 765), with the aim to predict their global quality of life (QoL) 18 months after the diagnosis and to identify possible QoL-related prognostic factors. For QoL prediction, a set of Machine Learning methods were explored, namely Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Depending on the model used, prediction accuracy varied between 0.305 and 0.864. Across models, a largely common set of psychological characteristics (optimism, perceived ability to deal with trauma, resilience as a trait, ability to understand the disease), as well as subjective perceptions of personal functionality (physical, social, cognitive function), were identified as key prognostic factors of long-term quality of life after a breast cancer diagnosis.**

Clinical Relevance—**Early detection of protective and obstructive factors associated with patient well-being can help health professionals to tailor preventive psychological programs aimed at enhancing the ability of breast cancer patients to adapt effectively to the disease.**

I. INTRODUCTION

In 2020, 2.3 million women worldwide were diagnosed with breast cancer and 685,000 died from the disease. By the end of 2020, there were 7.8 million 5-year breast cancer, making it the most prevalent cancer in the world [1]. Breast cancer occurs worldwide, especially in women over 30 years, with higher rates in middle age. Breast cancer is a significant and escalating global public health issue, evident in increasing incidence, mortality, and financial burden.

The diagnosis of breast cancer and subsequent medical treatments can often lead to significant psychological symptoms such as depression, anxiety, uncertainty, fear, loneliness, and body image problems [2]. Poor sense of well-being, as reflected in self-reported poor overall Quality of Life (QoL), is also frequently reported by patients throughout the critical period of breast cancer treatment [3]. Quality of life is acknowledged as an important, complementary, psychological outcome as well as a prognostic indicator of therapeutic outcome [4].

The problem of predicting the QoL of a breast cancer patient by a variety of factors is a field of active research. There are several studies that apply Machine Learning (ML) algorithms to medical data for classification, as well as algorithms for data mining, in order to find a pattern in the data set for faster calculations and prediction. For instance, using data from the present study, using only the RF-supervised learning algorithm, Mylona et al. [5] sought to classify breast cancer patients based on their mental health status at 6 months post-diagnosis and to identify potential prognostic factors. For the management of incomplete data, Mylona et al. [5] used the Multiple Imputation by Chained Equations (MICE) imputation method. The vast majority of important predictors were psychological. Model performance in terms of AUC (Area under the ROC Curve) ranged from 0.81± 0.04 to 0.90± 0.03. The cross-sectional study of Kourou et al. [6] applied a comprehensive supervised learning tool to a study population of 609 women with breast cancer discarding patients with a relatively high percentage of missing values, showing that personality and sociodemographic characteristics are strong predictors of health and health-related behavior. Kourou et al. [6] used the mean imputation to manage incomplete data. Best-performing approaches involved a meta-estimator combined with an SVM classification algorithm, exhibiting a balanced accuracy of 0.825.

The present study extends the prognosis of quality of life in patients with breast cancer to 18 months post-diagnosis. Importantly, we could include predictors measured during the psychologically critical period of cancer-related treatments (at 6 and/or 3 months post-diagnosis). However, in the present study, we include the whole 765 patients with breast cancer, a higher number compared to all the above-related studies, making it the most complete in terms of the data set. This study also uses both imputation methods, MICE and mean, with the most efficient one being the mean imputation.

In the present study, three models (RF, SVM, KNN) are examined, exploring multiple ML algorithms, imputation techniques, and oversampling methods, considering also data reaching up to six months after diagnosis. Further, feature selection in our approach was based on data items identified through deep literature research, besides traditional feature selection procedures.

The main goals of this study were (i) to assess the relative accuracy of several ML models in predicting poor global QoL of breast cancer patients at 18 months post-diagnosis utilizing a wide range of clinical, biological, psychological, socio-demographic and lifestyle data, and (ii) identify the

*This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777167 (BOUNCE).

C. R. is with the Hellenic Open University, E. M., K.K, G. M., H.K., K.M., P.S., E.K, D. I. is with FORTH, P.P. is with HUS, Ke. Ma is with IEO, R. P. is with HUJI, B. S. is with Champalimaud Foundation, (Corresponding author phone: +302810391449; e-mail: kondylak@ics.forth.gr).

most important prognostic factors that affect QoL. Alternative models differed on (a) the ML algorithm implemented (RF, SVM, and KNN), (b) the method used to handle missing values (replacement with global mean or Multiple Imputation through Chained Equations; MICE) [7], (c) the method employed to address class imbalance (none or Synthetic Minority Over-sampling Technique; SMOTE) [8] and (c) the set of predictor variables used for model cross-validation timing of measurement of predictor variables (data from M9 plus M3 vs data aggregated over the first 6 months peri- and post-diagnosis).

II. METHODS

A. Study Population

The study population consists of 765 women with breast cancer from a large multicenter study at five participating hospitals in four countries: (i) the European Institute of Oncology (IEO) in Italy (n = 213), (ii) the Rabin and Shaare Zedek Medical Centers (HUJI) in Israel (n = 151), (iii) the Helsinki University Hospital (HUS) in Finland (n = 238) and (iv) the Champalimaud Clinical Center (CHAMP)) in Portugal (n = 163). Inclusion criteria were age 40 to 70 years; a recent diagnosis of histologically confirmed invasive early or locally advanced operable breast cancer; tumor stage I to III; receiving surgery and any type of systemic treatment. Exclusion criteria included distant metastases and; a history of another malignancy within the last five years. Of the total cohort of 765 women enrolled, 495 (65.5%) were followed to 18 months averaging 55.5 years of age (SD=8.2).

Patient data included medical, socio-demographic, lifestyle, and psychosocial information assessed at seven-time points: at around the time of diagnosis (baseline) and at 3, 6, 9, 12, 15, and 18 months later. At baseline (M0) mostly non-cancer-specific measures were administered. As ontologies provide effective means to homogenize and integrate data [9], all the aforementioned data integrated using the BOUNCE psychological ontology [10].

B. Outcome Description

Global QoL assessed at 18 months post-baseline, using a single question on the EORTC Core Quality of Life (EORTC QLQ-C30) questionnaire, related to the patient's rating of overall QoL in the past week, served as the study outcome [11]. Individual scores on this measure were binarized with a cutoff value of 65/100 points corresponding to the lower quartile of the total sample distribution. Thus, a value of 0 indicated average and above average global QoL whereas a value of 1 indicated relatively poor QoL.

C. Predictor Variables

A large heterogeneous set of continuous and discrete variables was examined, including: a) clinical variables (e.g., medical history, BMI, stage of cancer, menopause, type of treatment); b) socio-demographic variables (e.g., age, education, marital status, income, employment); c) lifestyle variables (e.g., exercise, diet, smoking), and d) psychosocial variables from validated questionnaires (e.g., social support, mood optimism, sense of cohesion, coping flexibility, awareness, positive and negative impact, quality of life, etc.).

Each of the three types of models, imputation, and class-imbalance handling techniques was applied to six partially overlapping sets of predictor variables (see section D).

(i) The full set of 130 variables collected at M0 and M3

(ii) A subset of (i) containing the 67/130 most discriminating features (see Section D)

(iii) A subset from (i) including 30 hand-picked variables selected based on evidence from the literature ([12], [13], [14], [15], [16], [17], [18]).

(iv) The full set of the 186 variables collected at M0, M3, and M6

(v) A subset of (iv) including the 81 most discriminating features (see Section D)

(vi) A subset of (iv) including 35 hand-picked variables based on evidence from the literature ([12], [13], [14], [15], [16], [17], [18]).

D. Data Pre-processing and Feature Selection

All data were integrated and homogenized using an appropriate data infrastructure. Next, feature selection was implemented to (a) simplify the model by reducing the number of parameters, (b) reduce training time, (c) reduce overfitting and, (d) enhance generalization and reduce dimensionality. Of the total of 72 models tested in the present study data-driven feature selection was applied to one-third (14 models) using *permutation feature importance*. Permutation feature importance ranks the input features based on their importance in predicting the outcome variable.

Pre-processing of the data also included the imputation of missing values and over-sampling to take account of the class imbalance in the outcome variable. All dataset records include some missing data. The SMOTE technique [8] was applied to the minority class (poor QoL group) in order to create "synthetic" samples for the minority class, inserting new elements in the population.

For handling missing values, the MICE technique [7] was performed to maximize information and minimize bias due to incomplete data (10 imputations using 42 repetitions). In each repetition, there was an imputation for each attribute with missing values. The imputation methods included predictive mean matching for numerical variables, logistic regression for binary variables, and multivariate logistic regression for regular variables. Also, for the cases of missing values, another method for replacement was applied, replacing each missing value with the corresponding average value of each characteristic of the data set.

E. Model Building

For overall performance estimation the following 7 steps were followed:

Step 1. Data set preparation.

Step 2. To handle missing values two alternative approaches were applied: (i) the MICE imputation method, or (ii) the missing value (NaN) was replaced by the average of all the data of the respective characteristic categories.

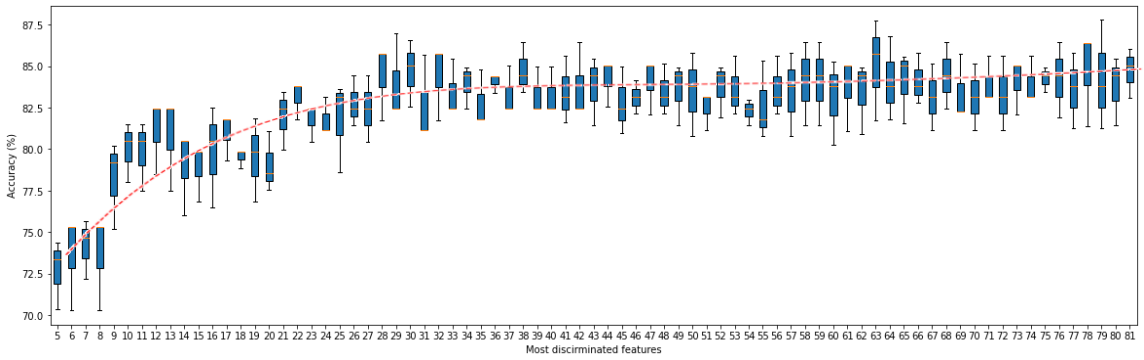

Figure 1. Mean balanced accuracy (SD in vertical bars) as a function of the number of features retained in the best-performing model (Model 40).

Thus, data imputation was performed after splitting the dataset into training and test, separately each, in order to avoid information leakage from the training to the test set. 20 percent of the dataset is allocated to the test set and 80 percent to the training set.

Step 3. For each case of Step 1, various dataset versions were created, using the most discriminating prognostic features, which affect the QoL of breast cancer patients and the ones with features identified from the literature as predictive or correlated with QoL.

Step 4. To avoid excessive adjustment and to make reliable predictions, a cross-validation (CV) scheme was followed, for the evaluation of the predictive models (RF, SVM, KNN), dividing the initial sample into a training set, to train the model and into a test set for its evaluation. A 10-fold CV was applied, where the original sample was randomly divided into 10 equal-sized models. Of the 10 folds, one is retained for validating the model and the remaining 9 are used as training data. Then, the CV process is repeated 10 times, so that each fold is used only once for validation.

Step 5. In each iteration, the oversampling method SMOTE was applied to the unbalanced training data set, in order to balance the two classifications.

Step 6. The RF, SVM, and KNN classifiers were selected for model training. The balanced set was further divided into 10 sections and CV was repeated 10 times to optimize hyperparameters. At the end of the process, the prediction accuracy of the classifier was calculated.

Step 7. The performance of each model was evaluated in the corresponding test set, using balanced accuracy. The overall performance of each model was calculated as the average performance in each test set. Prediction accuracy was used as a performance measurement and the best model was selected based on this measurement.

I. RESULTS

Model performance accuracy. A total of 72 models were run and Table I shows the models with the 10 highest overall prediction performance (balanced accuracy). Model 40 demonstrated the highest balanced accuracy (0.864). This model applied the RF algorithm on an optimized set of 81 variables from M0, M3, and M6, data imputation using mean replacement, and SMOTE to address the class imbalance. In general, RF models outperformed SVM and KNN models.

Running this model, using the 81 most important features, starting with the 5 most important and incrementing them one by one in descending order of importance, we can identify that after 30 features the performance is not notably improved (see Figure 1). The most discriminating features were selected based on their positive impact (weight) on the QoL of breast cancer patients.

Predictor variables. Based on all three methods (RF, SVM, and KNN), consistent variables were identified, which indicate both psychological (primarily) and physical characteristics, as the most important contributions to the QoL classification of breast cancer patients. The 10 most important parameters identified, based on their impact, are the following:

I. Symptoms of post-traumatic stress disorder (PTSD) 6 months after diagnosis, which develops in some people who have experienced a shocking, frightening, or dangerous event such as breast cancer [20], where the impact (weight) ranges between 0.3682 ± 0.1308

II. & IV. The number of sick leave days taken by the patient during the first 3 (weight: 0.0545 ± 0.0438) and 6 months (weight: 0.0099 ± 0.0159) after diagnosis.

III. Self-efficacy 6 months after diagnosis (weight: 0.0278 ± 0.0365), a patient's belief that she can adapt the behavior needed to produce the expected outcome concerning cancer's consequences and treatment.

V. The severity of pain experienced at 6 months after diagnosis (weight: 0.0088 ± 0.0282). Pain is one of the most common symptoms in cancer patients.

VI. Fear of cancer recurrence, 6 months after diagnosis (weight: 0.0073 ± 0.0073). Most people who have been diagnosed with cancer, worry about recurrence i.e., the possibility that the cancer will return or progress to the same or another part of the body [20].

VII. & X. Diarrhea that occurs at baseline (weight: 0.0064 ± 0.0043) and 6 months later (weight: 0.0060 ± 0.0167). Some cancer treatments can cause diarrhea, including chemotherapy, radiation, surgery, and bone marrow transplants.

VIII. Role functioning scale at baseline (weight: 0.0062 ± 0.0076), refers to the patient's ability to perform daily activities, leisure activities, and/or work.

IX. Trait resilience (measured at baseline) (weight: 0.0061 ± 0.0061). The concept of resilience includes the protective qualities and/or personal characteristics of a person, which help in the successful adaptation to cancer.

TABLE I. MODELS WITH THE HIGHEST ACCURACY

#	*Data*	*Imputation*	*Over sampling*	*Method*	*Balanced Accuracy*	*sensitivity/ specificity*
40	m0, m3 & m6	Mean	SMOTE	RF	**0.864**	0.934/ 0.606
38	m0, m3 & m6	MICE	SMOTE	RF	0.857	0.915/ 0.729
27	m0 & m3 – liter. based	Mean	without SMOTE	RF	0.85	0.917/ 0.394
63	m0, m3 & m6 – liter. based	Mean	without SMOTE	RF	0.845	0.942/ 0.424
39	m0, m3 & m6	Mean	without SMOTE	RF	0.84	0.95/ 0.545
26	m0 & m3 – liter. based	MICE	SMOTE	RF	0.831	0.893/ 0.667
52	m0, m3 & m6 – most discriminating features	Mean	SMOTE	RF	0.831	0.959/ 0.515
37	m0, m3 & m6	MICE	without SMOTE	RF	0.825	0.943/ 0.667
50	m0, m3 & m6 – most discriminating features	MICE	SMOTE	RF	0.825	0.915/ 0.625
64	m0, m3 & m6 – liter. based	Mean	SMOTE	RF	0.825	0.926/ 0.455

II. CONCLUSION

By selecting different datasets, machine learning methods, and preprocessing steps, 72 models were developed to identify the factors that lead to the highest prediction accuracy. Models 1 to 36 aimed to predict QoL status from data at M0 and M3, while models 37 to 72 aimed to predict QoL status from data at M0, M3, and M6. Each of these two classes of models used a combination of the three machine learning methods RF, SVM, and KNN, data imputation with MICE and replacement with the mean, as well as estimated accuracy with and without the application of SMOTE data oversampling. The best prediction performance was observed in Model 40, where the RF method was applied, which used M0, M3, and M6 data, a method of replacing missing values with the mean was used for data imputation, and the SMOTE over-sampling technique was applied. For the given dataset, it is observed that the RF method has the best performance compared to the SVM and KNN. Also, it is observed that for RF and SVM methods, the more data available, the more efficient the models are. Specifically, the best performance is obtained when the dataset consists of all data at M0, M3, and M6 (186 features). As data are removed, either by reducing the time range from which data are used i.e., for M0 and M3, or by using the most significant features (from M0, M3, and M6), or by using the features that are reported as important for predicting QoL in breast cancer patients in literature reports and research, their performance decreases. Conversely, the more data the KNN method includes, the less efficient it is. Thus, the lowest performance of all the models is obtained by Model 48, with an accuracy of 0.305, which uses the KNN method and includes the dataset from time points M0, M3, and M6 (186 features), as well as apply data imputation by applying the mean and uses the SMOTE over-sampling technique. In contrast, KNN performs best when using fewer data (81 features), i.e., it uses the most relevant features from the dataset from time moments M0 and M3. In conclusion, we can say that the more data before the 18-month time interval for which we want to find the prediction accuracy, the higher the performance of the RF and SVM methods.

REFERENCES

[1] E. Jutta, et al., "Predictors of Quality of Life of Breast Cancer Patients", *Acta Oncologica*, 2003, Vol. 42 (Issue 7), pp. 710-718.

[2] S. Xu, et al., "The Global, Regional, and National Burden and Trends of Breast Cancer From 1990 to 2019: Results from the Global Burden of Disease Study 2019". *Front Oncol*, 2021.

[3] A. Elisabeth, et al., "Global quality of life and its potential predictors in breast cancer patients: an exploratory study". *Supportive Care in* Cancer, 2007, 15.1, pp. 21-30.

[4] P. Theofilou, "Quality of life outcomes in patients with breast cancer". *Oncology Reviews*, 2012, Vol. 6(1).

[5] E. Mylona, et al., "Prediction of Poor Mental Health Following Breast Cancer Diagnosis Using Random Forests", *EMBS*, 2021.

[6] K. Kourou, et al., "Computational models for predicting resilience levels of women with breast cancer", *IFMBE Proceedings*, 2020.

[7] M.J. Azur, et al. "Multiple imputation by chained equations: what is it and how does it work?", *IJMPR*. 2011, Vol. 20(1): pp. 40–49.

[8] S. Satpathy, "Overcoming Class Imbalance using SMOTE Techniques" *Data Science Blogathon*. 2020.

[9] L. Martin, et al., "Ontology Based Integration of Distributed and Heterogeneous Data Sources in ACGT". *HEALTHINF*, 2008, 301-306.

[10] H. Kondylakis, et al., "Developing the BOUNCE Psychological Ontology" *ISWC*, 2020.

[11] H. Knobel, "The validity of EORTC QLQ-C30 fatigue scale in advanced cancer patients and cancer survivors". *Palliat Med*, 2003, Vol. 17(8): pp. 664-72.

[12] N. Sharma, A. Purkayastha, "Factors affecting Quality of Life in Breast Cancer Patients: A Descriptive and Cross-sectional Study with Review of Literature", *J Midlife Health*. 2017;8(2):75-83.

[13] C. Fincka, et al., "Quality of life in breast cancer patients: Associations with optimism and social support", *International Journal of Clinical and Health Psychology*, 2018, Vol. 18. Issue 1., pp. 27-34.

[14] M.R. Zaker, A. Hazrati-Marangaloo, S.R Hosseini, "Quality of Life in Iranian Breast Cancer Survivors and Affecting Factors: A Review Article", *MSc in Nursing*, Urmia University of Medical Science.

[15] A. Safaee, "Predictors of quality of life in breast cancer patients under chemotherapy", *Indian Journal of Cancer*, 2008, Vol 45(3).

[16] H. Michelson, "Health-related Quality of Life Measured by the EORTC QLQ-C30: Reference Values From a Large Sample of the Swedish Population", *Acta Oncologica*, 2000, 39:4, 477-484.

[17] H.Y. Huang, et al., "Quality of life of breast and cervical cancer survivors", *BMC Women's Health*, 2017, 17:30.

[18] N. Sharma, A. Purkayastha, "Factors affecting quality of life in breast cancer patients: A descriptive and cross-sectional study with review of literature", *Journal of Mid-Life Health*, 2017, Vol 8(2).

[19] H. Kondylakis, et al., "Developing a data infrastructure for enabling breast cancer women to BOUNCE back", *CBMS*, 2019, pp. 652–657.

[20] S. Swartzman, et al., "Posttraumatic stress disorder after cancer diagnosis in adults: A meta-analysis". *Depression and anxiety*, 2017, 34(4), 327–339.