Multimedia Tools and Applications (2021) 80:13787–13800
https://doi.org/10.1007/s11042-020-10448-w

Predicting the recurrence of breast cancer using machine learning algorithms

Amal Alzu'bi[1] · Hassan Najadat[1] · Wesam Doulat[2] · Osama Al-Shari[3] · Leming Zhou[4]

Received: 26 January 2020 / Revised: 24 August 2020 / Accepted: 22 December 2020 /
Published online: 18 January 2021

Abstract
Breast cancer is one of the most common types of cancer among Jordanian women. Recently, healthcare organizations in Jordan have adopted electronic health records, which makes it feasible for researchers to access huge amounts of medical records. The goal of this study is to predict the recurrence of breast cancer using machine learning algorithms. We developed a Natural Language Processing algorithm to extract key features about breast cancer from medical records at King Abdullah University Hospital (KAUH) in Jordan. We integrated these features and built a medical dictionary for breast cancer. We applied multiple machine learning algorithms on the extracted information to predict the recurrence of breast cancer in patients. Our predicted results were approved by specialist physicians from KAUH. The medical dictionary was created and the accuracy of the data had been validated by targeted users (physicians, researchers). This dictionary can be used for personalized medicine. All machine learning algorithms had a nice performance. OneR algorithm has the best balance of sensitivity and specificity. The medical dictionary will help physicians to choose the most appropriate treatment plan in a short time. The machine learning prediction results can help physicians to make the correct clinical decision regarding their treatment options.

Keywords Machine learning · Natural language processing · Healthcare · Breast cancer

1 Introduction

Breast cancer is one of the most common types of cancer and one of the leading causes of death among Jordanian women [1]. Based on the report from the Jordan Cancer Registry (JCR) 2011, breast cancer represents 20.1% of overall cancer cases for both males and females in Jordan, and 37.3% of overall cancer cases in females.

✉ Amal Alzu'bi
aazoubi9@just.edu.jo

Extended author information available on the last page of the article

Recently, healthcare organizations in Jordan have adopted electronic health record (EHR) systems to improve the efficiency of their work [6]. EHRs include informative descriptions about patients' diagnosis, symptoms, medications, and doctors' medical recommendations, which can be highly valuable for medical research.

However, it is difficult to extract accurate and useful information from the unstructured clinical documents without extensive interpretation from physicians [22]. The active involvement of physicians in research studies is not feasible for many researchers. Therefore, there is a need to transform the unstructured medical records from free-text reports into a structured format using a computer algorithm.

Natural Language Processing (NLP) algorithms can analyze, process and interpret documents written in natural language based on a set of theories and computational methods [40]. Many studies have already employed NLP on EHR data. For instance, an automated tool was developed for extracting medical problems from EHR [31]. A hybrid approach of NLP and International Classification of Disease-9th Revision (ICD-9) codes was used to identify hepatocellular cancer from EHR data [33]. Therefore, NLP appears to be the solution to extract the needed information from the unstructured and heterogeneous medical data, for instance, medical notes and radiology image reports [40].

In medicine practice, physicians need to integrate all available patient data including, symptoms, family history, medications, allergy, lab results and genomic information in order to have the most accurate diagnosis and thus determine treatment options for each patient. [7]. In some cases, physicians may need to test several different treatment plans before they find the best treatment option for their patients [25]. In some types of cancer, however, the disease can change from one stage to another in a short period of time and becomes harder to treat [23]. Therefore, it is desired to identify the optimal treatment approach quickly. Machine learning may provide the desired solution.

Machine learning algorithms have been applied in many fields, such as topography [28], energy management [26, 42], and text document classification [2, 4, 3], and preventive maintenance [24]. The results from these machine learning applications help us to make optimal decision. Below is a brief summary of some machine learning applications in disease prediction.

Machine learning algorithms can be widely used in the field of medical analysis. For example, Machine learning algorithms can be used for early disease recognition and prediction [39, 16], such as the prediction of infectious diseases [13], chronic diseases [11], pulmonary hypertension disease [17], liver diseases [22] and heart disease [34] and cancer disease (9).

Cancer prediction is one of the applications of machine learning algorithms. There are several studies that applied machine learning algorithms to predict and determine the recurrence of breast cancer disease. For example, authors in [5] used three machine learning algorithms, including decision tree (DT), artificial neural networks (ANN), and Support Vector Machine (SVM). They studied 22 features of breast cancer and their result showed that SVM has the highest accuracy and the minimum error rate. Authors in [38] used different machine learning algorithms, including random forest, SVM, logistic regression, and bayesian classification algorithm to build a prediction model based on three biomarkers. The model could successfully predict the recurrence of the disease in patients in very early stages. Researches in [12] proposed a model for predicting breast cancer recurrence using SVM and ANN. ANN could predict the recurrence of the disease with high accuracy.

Different risk factors and predictors of the recurrence chance of breast cancer have been discussed in the literature. For example, authors in [14] used ensemble learning algorithm called XGBoost to predict the recurrence of breast cancer using 23 predictors. Machine learning algorithms help physicians to identify the responsible risk factors for breast cancer recurrence. These risk factors include: type of cancer therapy [18], molecular data and susceptible genes [41], body mass index (BMI) [20] and molecular subtype [35] .

Although several studies discussed the risk factors and predictors of the recurrence chance of breast cancer, it is still an open field for research. Machine learning algorithms are sensitive to the input data and thus there is no one suitable algorithm for all data types. In Our study, we analyzed the available data at KAUH. We studied several risk factors and predictors and applied multiple machine learning algorithms in order to find the most accurate algorithm and parameters to predict the recurrence chance of breast cancer in patients at KAUH. We validated the results using different measures including, accuracy, sensitivity, specificity, and error rate.

In our study, we aim to build a structured medical dictionary for breast cancer through extracting and integrating all related information from multiple sources in order to provide a comprehensive database about the disease. We used the data in the dictionary to build a machine learning model in order to predict the recurrence of the breast cancer in patients and thus make the correct decision regarding their treatment options and optimized therapies. Early prediction of the recurrence of breast cancer will save the life of patients. Additionally, our study will give an overview about the main biomarkers' predictors of breast cancer among Jordanian women.

The rest of the paper is organized as follows. Section 2 describes the methodology of our work. Section 3 presents the experiments result. Section 4 provides a discussion and finally in Section 5, we provide our conclusion.

2 Methods and materials

2.1 General procedure

As mentioned in the introduction, the study has two major components. The first one is an NLP algorithm for extracting important features about breast cancer from EHR. The NLP algorithm was created using Python script. This algorithm scanned through diagnosis, symptoms, medications, lab results, recommendations, past medical history, procedures, family history, imaging, endoscopic assessment, anesthesia type, allergies, and others clinical documents in the EHR at King Abdullah University Hospital and searched for some important features related to breast cancer. The extracted information was used to build the structured medical library for breast cancer.

In the second component, we used machine learning algorithms to build a model for predicting the recurrence chance of the breast cancer in patients, which can be used to guide physicians to make informed decision regarding the treatment options and optimized therapies. Figure 1 demonstrates the architecture of our work.

2.2 Data collection and preprocessing

A total of 1475 patient records from King Abdullah University Hospital were collected. First, we deleted the duplicated records and excluded the records with several missing values and the records for patients with other diseases such as prostate cancer. Second, we selected the cases that have standard histological examination reports, including results of hormone receptors (ER, PR, HER-2) and tumor specification.

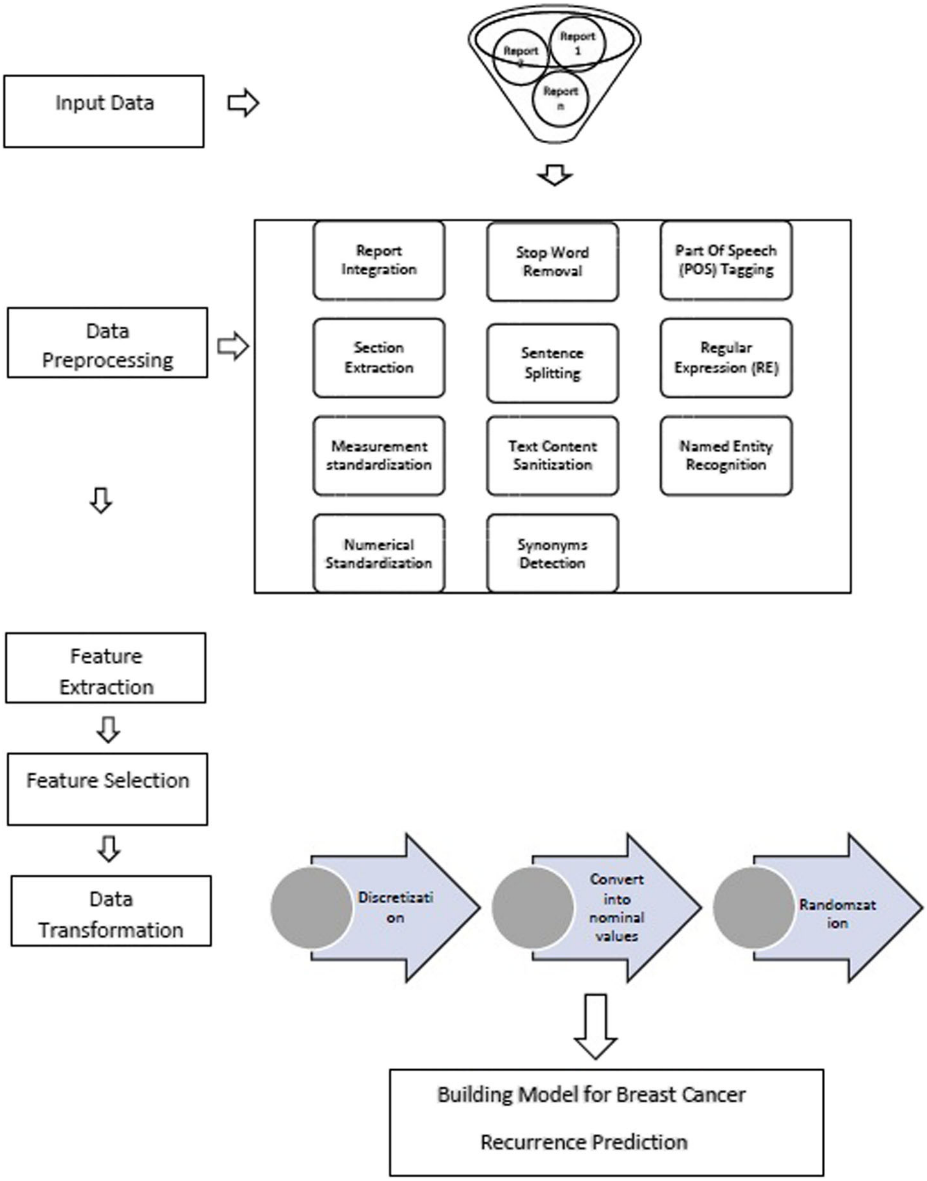

Fig. 1 System architecture

Third, we used the TNM (T stands for tumor size, N describes the lymph nodes, and M describes metastasis or the spread of cancer) staging system [9] and the 8th edition of American Joint Committee on Cancer (AJCC 8th) for anatomic staging [8]. We put breast cancer patients into different categories according to their molecular subtypes: Luminal A (LA), Luminal B (LB), HER-2 positive, and Triple Negative Breast Cancer (TNBC) [19].

Fourth, we identified 142 breast cancer histopathology reports written in free text. Every report has seven sections. The first section is about clinical history and differential diagnosis section. The second section is about tissue source, which indicates the examined mass side. The third section is about specimens, which describes the breast abnormality and any underwent surgeries history. The fourth section is the gross description, which gives information about the specimen size, biopsies measurement and weight. The fifth section is about the diagnosis, which includes information about tumor size, type of cancer, tumor grade, hormone receptors, and test results. The six section provides a summary, which summarizes the important test result from the examined tissues. The last section provides some comments and notes from the laboratory about the requested tests.

In the next step, we used the following techniques to preprocess the collected data:

- Report Integration, which integrates multiple pathology reports related to the same patient into one report based on the patientID.
- Section Extraction, which extracts the content of each section separately.
- Measurement standardization, which converts all measurement into one standard measure unit. For example, greatest dimension of tumor size that is found in centimeters (cm) and millimeters (mm). We chose to use mm as the unit for all tumor size.
- Synonyms detection: Many synonyms are detected and approved by physicians.
- Text content sanitization: in this step, we removed noisy contents which are not important for the description of the disease, such as symbols and dates. We then omitted any content that does not represent medical information using the Natural Language Toolkit (NLTK). e.g. (-,),(, <, >, /, [,], *, :)
- Sentence splitting: we used the NLTK to detect and split the reports into sentences.
- Stop word removal: we removed all frequently used stop words such as, in, on, the, etc.
- Part of speech (POS) Tagging: in this step, NLTK POS tagger was used to define any part of the speech tags such as, noun, verb, determinant.
- Regular expression (RE): we used Python regular expressions to specify the set of features that are used in our dataset. Table 1 shows the extracted features from the EHR.
- Named entity recognition: in this step, we used the NLTK wrapper library to extract the words that express named entities such as person and location.

The results of all these steps were thoroughly evaluated by the breast cancer specialists to ensure their correctness.

2.3 Feature extraction

To find the importance of each term in the reports, we used the Term Frequency - Inverse Document Frequency (TF-IDF) method. TF/IDF returns the most important keywords that can describe a feature, symptom, diagnosis, or any important terms. The term frequency of the word in each report within the EHR is a ratio of the number of times the term appears in a report and the total number of words in the report.

$$tf_{ij} = \frac{n_{ij}}{\sum n_{ij}}$$

here, tf_{ij} is the frequency of term i in report j, n_{ij} is the number of times the term i shows up in report j that contain the term i, N is the total number of reports.

The Inverse Document Frequency is used to measure how common each word appears in all reports.

$$idf(i) = log\frac{N}{df(i)}$$

The TF-IDF of the term i at report j is:

$$w_{ij} = tf_{ij}log\frac{N}{df(i)}$$

Where $df(i)$ is the number of reports that contain the term i, and N is the total number of reports.

We processed the text within the reports using the TF-IDF to determine the important keywords that can describe the meaningful features, such as symptoms, diagnoses, and all other important terms.

Before building a predictive model, we need to perform the step of feature selection. First, we identified all important features of breast cancer available in the literature by performing a comprehensive literature review. Table 1 lists the key breast cancer factors that we identified from the literature.

After reviewing the import breast cancer features or risk factors in the literature, we analyzed 1475 clinical documents at KAUH to search for the identified features. We then asked the specialist physicians to rank the identified features and determine the features that are related to our research. Additionally, to perform the feature selection, we used the InfoGainAttributeEval method, which determines the value of the attribute by measuring the information gain with respect to the class (recurrence class). It calculates the amount of mutual information gained from the feature with respect to the class (recurrence class). We selected the features with a cutoff value 0.05. Table 2 represents our extracted features with their possible values.

Moreover, we added some additional features to the dictionary such as, the values of T, N, M, ER, PR, HER-2, which are important to determine the anatomic stage AJCC and the molecular subtype.

Table 1 Features identified from the literature

Authors	Year	Predictors of breast cancer
Chung et al. [15]	2019	LVI, PR, CK5/6
Bakre et al. [10]	2019	Bio markers (CD44, ABCC4, ABCC11, N-Cadherin, and Pan-Cadherin), and TNM
Lafourcade [29]	2018	Tumor grade, tumor size, axillary nodal involvement, ER, PR
Zhou et al. [43]	2016	lncRNA biomarkers
Song et al. [36]	2012	Menopausal status, operation method, stage, nodal status, histologic grade, nuclear grade, extensive intraductal carcinoma component, hormone receptor, p53, c-erbB-2, Ki-67, and molecular subtype.
Filipits et al. [21]	2011	ER, HER-2
Partridge et al. [32]	2005	Tumor volume and diameters
Meric et al. [30]	2003	Age, tumor size, tumor grade, estrogen and progesterone receptor status, surgical margins, axillary lymph node involvement, and use of adjuvant therapy.
Huang et al. [27]	2003	Gene expression predictors of breast cancer outcomes
Björn et al. [37]	1982	Axillary metastization, size of primary tumor, and differentiation of the primary tumor

Table 2 Our extracted features and their possible values

Feature	Description	Possible values
Age	Patient age	Numeric values
Menopause	Menopause status	Menopause, premenopausal
DCIS	DCIS types	Comedo, solid, cribriform, micropapillary
LVI	Lympho Vascular Invasion	Present, not present
T	Tumor size	T1, T2, T3, T4
N	Number of lymph node	N0, N1, N2, N3
M	Metastasis	M0, M1
PR	Progesterone receptor	Positive, negative
ER	Estrogen receptor	Positive, negative
HER-2	Human epidermal growth factor receptor 2	Positive, negative
GD	Greatest dimension of primary tumor	Numeric values
Focality	Cancer focality	Unifocal, multifocal
Target therapy	Cancer therapy	LETROZOLE (2.5 MG) EXEMESTANE (25 MG) TAMOXIFEN (20 MG)
Mastectomy	Mastectomy surgery	Yes, no

2.4 Data transformation

Data transformation is an important step in Machine Learning. In our work, we used multiple transformation processes in order to make sure that the dataset is ready to be used in machine learning algorithms. These transformation processes include:

- Discretization, which converts continuous numeric values into small groups. In our dataset, we used this on the age and the dimension of the primary tumor. The age is organized into three groups: < 44, between 45 and 65, and > 65. Groups were selected using discretize filter in WEKA.
- Converting numeric values into nominal. In WEKA, we used the Numeric to Nominal filter to convert all numeric values into nominal values, such as molecular subtype and AJCC anatomic stage, in which the molecular subtypes are classified into four groups (1, 2, 3, 4). 1 stands for LA, 2 stands for LB, 3 stands for HER-2 positive and 4 stands for TNBC. In this filter, WEKA considers each value as a nominal value to get accurate classification results.

2.5 Building the medical dictionary of breast cancer

In this step, we built a medical dictionary of breast cancer based on the extracted features in the previous step. We used Python scripts to save these features into a database. The dictionary contains detailed information about breast cancer, including, patient's age, diagnosis, clinical tests, DCIS subtypes, LVI, tumor characteristics, TNM staging, cancer focality, hormonal receptors status and the cancer therapies that are used in the treatment plan.

To make the dictionary comprehensive and informative, we added information about the anatomic staging and the molecular subtype of breast cancer. Defining the anatomic stage and molecular subtype provides a convenient way to describe the cancer that helps physicians to decide the best cancer therapy, surgeries intervention, and treatment plan. The anatomic

Table 3 Characteristics of molecular subtypes according to the size of the tumor

	LA	LB	HER-2 positive	TNBC
T1	5	11	2	5
T2	22	34	8	10
T3	6	16	8	4
T4	1	5	2	2

staging of breast cancer is based on the AJCC staging manual and based on the TNM and hormone receptors status. The molecular subtypes are based on the hormone receptors status.

The dictionary also contains detailed information about the risk factors of breast cancer including, cancer stage, molecular subtype, diagnosis, test type, and cancer therapies.

2.6 Building the predictive models

We used machine learning algorithms to predict the factors that increase the risk of recurrence and determine whether the patient needs different cancer therapy or additional therapies. Therefore, we used Weka 3.9 toolkit to perform our experiments using nine classifiers. The performance of each classifier was validated and measured based on the accuracy, sensitivity, specificity, time to build the model, and the error rate. In our experiment, we used 9 classifiers and 10-fold cross validation to make a comparison between the classifiers. In this method, each one of the 10 subsets works as an independent set to train the model on the rest of subsets. Each fold contains a pair of training and testing sets.

We evaluated the extracted features including, tumor grade, molecular subtype, cancer focality, LVI, menopause, DCIS type, age, and greatest dimension of primary tumor as predictors of breast cancer recurrence chance using different machine learning algorithms including, J48, NaïveBayes, bagging, logistic, SVM, KNN, MLP, PART, and OneR.

3 Results

3.1 Descriptive statistics

In this study, there were 142 cases of ductal carcinoma in situ (DCIS), a type of non-invasive breast cancer. These 142 patients were categorized into four DCIS subtypes: Comedo, Solid, Cribriform, and Micropalillary. Out of these 142 female patients, 43 (30.3%) were at age 50 years old; 60 (42.3%) were younger than 50 s, and 39 (27.4%) were older than 50 years old.

Among these 142 cases, 46% of them were Luminal B, 24% were HER-2 positive, 15% were TNBC, 15% were LA.

Table 4 Characteristics of molecular subtypes according to lymph nodes involvement

	LA	LB	HER-2 positive	TNBC	Total
N0	6	12	2	10	30
N1	10	26	5	6	47
N2	6	19	6	4	35
N3	12	7	7	1	

Table 5 Characteristics of molecular subtypes according to tumor grade

	LA	LB	Her-2 positive	TNBC	Total
G1	2	13	0	2	17
G2	10	32	1	1	44
G3	22	19	20	16	77
G4	0	2	0	2	4
Number	34	66	21	21	142

In all these cases, 40% were diagnosed at pre-menopause ages (< 45 years) and 60% of all cases were diagnosed after menopause age (> 45 years old).

The data showed that most of the cases are T2 according to TNM staging, which means that the tumor size is between 2 cm and 5 cm (centimeters). Table 3 represents the characteristics of the molecular subtypes according to lymph nodes involvement.

Tumor grade provides important details about the characteristic of the tumor under microscope. In LA, HER-2 positive, and TNBC, most cases have G3 tumor grade, which means that the cancer cells are poorly differentiated and thus the tumor grows rapidly. In LB, tumor grade is mostly G2 tumor grade, which means that the tumor tends to grow slowly, and the tumor cells are well differentiated. Table 4 summarizes these results. Defining the characteristics of the molecular subtypes according to the anatomic cancer stage is shown in Table 5. 30% of LA cases were in stage IA. In LB cases, 30% of the cases were at stage IIIB and 20% were at stage IA. Most of the HER-2 positive cases were in IIB stage. In TNBC cases, 71% of cases were in stage IIIC.

Figure 2 represents the distribution of our cases within different anatomic cancer stages. We can notice that most of our cases are at menopause age.

3.2 NLP results

We identified 307 terms with non-zero values in TF-IDF from those clinical reports. The higher values of TF-IDF indicate those infrequent but more distinguishable terms. For example, the highest value of TF-IDF corresponds to “neck mass” (TF-IDF = 0.1602), which

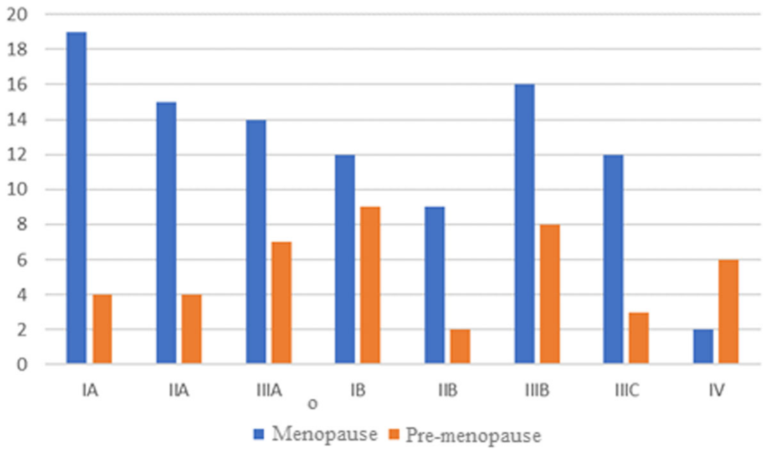

Fig. 2 Anatomic cancer stages

Table 6 Our selected features

Feature	Rank
Tumor grade	0.30282
AJCC	0.29085
Molecular subtype	0.27394
Cancer focality	0.23732
LVI	0.17676
Menopause	0.14296
DCIS type	0.14366
Age	0.20282
Greatest dimension of primary tumor	0.10258

is unique for some breast cancer patients. While the lowest value of TF-IDF is for "carcinoma" (TF-IDF = 0.0080), which is not unique for cancer related reports.

All these results were integrated into a comprehensive database. The database includes information about patients age, gender, diagnosis, treatment type, cancer therapy, test name, cancer foci, breast mass side, DCIS type, menopause status, tumor size, dimension of the primary tumor, number of lymph nodes that have a cancer, metastasis status, HER-2 result, ER result, PR result, molecular subtype, anatomic stage, survivalist, and the recurrence status.

3.3 Classification results

In order to build the model that aims to predict the recurrence chance of breast cancer disease, we used InfoGainAttributeEval method with a cutoff value of 0.05 to select the features that are relevant to breast cancer recurrence chance. Table 6 shows our selected features.

As shown in Fig. 3, bagging classifier provides high sensitivity value and low specificity. This means that bagging classifier can capture most of the recurrence chance of breast cancer, but also have many false positives. OneR and J48 classifiers have high sensitivity and specificity values, which means that these classifiers are able to identify women who have the chances of recurrence. Additionally, they are able to correctly identify women who need to

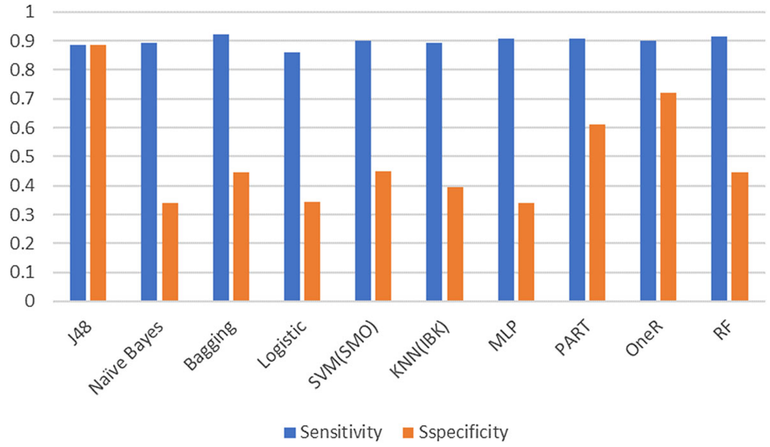

Fig. 3 Sensitivity and specificity of the classifiers

test for recurrence. Although OneR and J48 classifiers have very similar results in terms of sensitivity, specificity, and time, OneR classifier has the smallest error rate.

Table 7 summarizes the accuracy, sensitivity, specificity, error rate and time values of our classifiers.

Different classifiers need different time to build the model. J48, NaïveBayes, PART, and OneR need short time, while MLP classifier needs the maximum time to build the model. Although Bagging algorithm has the highest accuracy, OneR classifier doesn't need time to build the model and has the minimum error rate.

In order to compare the time needed to extract the main features of breast cancer using our algorithm with the time needed to extract the same features manually, we asked the specialist physicians to manually extract the main features of breast cancer. We found that the time needed to extract these features from different resources and save the result into a useful dataset using our algorithm is just a few seconds, while the time needed for manual extraction of the same features is roughly (4–5) minutes.

4 Discussion

Based on our analysis, we could find that the most critical factors that can predict the recurrence possibility in breast cancer patients including; tumor grade, molecular subtype, cancer focality, LVI, menopause, DCIS type, age, and greatest dimension of primary tumor.

The information provided in our database can create opportunities to improve patients' care by providing some recommendations about patients' treatment options. Additionally, it will help physicians to choose the most appropriate treatment plan in short time based on the matching between their patients and the patients available in the dictionary. It will also help physicians to predict the molecular subtype and the stage of the disease. Furthermore, it creates very big opportunity to improve research since it provides a comprehensive database that integrates data from multiple sources and makes them available at one place. Therefore, researchers can easily query for any set of variables/features and study the relationship between them within a group of patients.

One limitation of our approach is that it mainly uses clinical data from King Abdullah University Hospital (KAUH). One more limitation is the unstructured and variable format of the clinical data stored in the hospital EHR, which increases the complexity and variability of our extraction algorithm. Additionally, several cases were lost in the follow up and several medical records have missing values, which enforced us to exclude these records.

Table 7 Different classifiers and their reported measurement values

Classifier	Accuracy	Sensitivity	Specificity	Error rate	Time
J48	88.73	0.887	0.887	0.2002	0.0
Naïve Bayes	89.44	0.894	0.341	0.1414	0.0
Bagging	92.25	0.923	0.446	0.1474	0.23
Logistic	85.92	0.859	0.345	0.1661	0.04
SVM(SMO)	90.14	0.901	0.449	0.0986	0.02
KNN(IBK)	89.44	0.894	0.395	0.1453	0.0
MLP	90.85	0.908	0.339	0.1063	0.44
PART	89.44	0.908	0.612	0.1228	0.0
OneR	90.1408	0.901	0.722	0.0986	0.0

5 Conclusions

In this work, we used NLP and machine learning algorithms to extract beneficial features of breast cancer form unstructured EHR data and store them into a comprehensive database, then we built a prediction model that can predict the recurrence chance of the breast cancer in patients and thus make the correct clinical decision regarding their treatment options and therapy. Our Results showed that bagging classifier outperforms the other classifiers used in the experiments. OneR has the minimum error rate and the minimum time to build the model. It also has the best specificity and sensitivity. We positively expect that our model has the ability to be applied within larger datasets and to achieve high accuracy.

References

1. Abdel-Razeq H, Attiga F, Mansour A (2015) Cancer care in Jordan. Hematol Oncol Stem Cell Ther 8(2): 64–70
2. Abualigah L (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Studies in computational intelligence. Springer International Publishing, Berlin
3. Abualigah L (2020) Multi-verse optimizer algorithm: a comprehensive survey of its results, variants, and applications. Neural Comput Applic 32:12381–12401
4. Abualigah L, Khader A (2017) Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J Supercomput 73:4773–4795. https://doi.org/10.1007/s11227-017-2046-2
5. Ahmad L, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A (2013) Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform 4(2). https://doi.org/10.4172/2157-7420.1000124
6. Al-Adwan A, Berger H (2015) Exploring physicians' behavioural intention toward the toward the adoption of electronic health records. Int J Healthc Technol. Manag 15(2):89–111
7. Alzu'bi A, Zhou L, Watzlaf V (2014) Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals. Perspect Health Inf Manag 11(Spring):1c
8. Amin M et al (2017) The eighth edition ajcc cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 67(2):93–99
9. Bagaria S et al (2014) Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. JAMA Surg 149(2):125–9
10. Bakre M et al (2019) Clinical validation of an immunohistochemistry-based canassist-breast test for distant recurrence prediction in hormone receptor-positive breast cancer patients. Cancer Med 8(4):1755–1764
11. Battineni G et al (2020) Applications of machine learning predictive models in the chronic disease diagnosis. J Perinat Med 10(2):21
12. Boeri C et al (2020) Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med 9(9):3234–3243
13. Chae S, Kwon S, Lee D (2018) Predicting infectious disease using deep learning and big data. Int J Environ Res Public Health 15(8):1596
14. Chang C, Chen S (2019) Developing a novel machine learning-based classification scheme for predicting spcs in breast cancer survivors. Front Genet 10(848). https://doi.org/10.3389/fgene.2019.00848
15. Chung S et al (2019) Prognostic factors predicting recurrence in in- vasive breast cancer: An analysis of radiological and clinicopathological factors. Asian J Surg 42(5):613–620
16. Dahiwade D, Patle G, Meshram E (2019) Designing disease prediction model using machine learning approach, in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, New York
17. Dawes T et al (2017) Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. Radiology 283(2):381–390
18. Eidemüller M et al (2019) Long-term health risk after breast-cancer radiotherapy: overview of passos methodology and software. Radiat Prot Dosim 183:259–263
19. Falck A, Fernö M, Bendahl P, Rydén L (2013) St Gallen molecular subtypes in primary breast cancer and matched lymph node metastases–aspects on distribution and prognosis for patients with luminal A tumours:

results from a prospective randomised trial. BMC Cancer 13(558). https://doi.org/10.1186/1471-2407-13-558

20. Feliciano E et al (2017) Body mass index, pam50 subtype, recurrence, and survival among patients with nonmetastatic breast cancer. Cancer 123(13):2535–2542
21. Filipits M et al (2011) A new molecular predictor of distant recurrence in er-positive, her2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res 17(18):6012–6020
22. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA (2016) Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 23(5):1007–1015
23. Gerhard W. The diagnosis, pathology, and treatment of the diseases of the chestchest. Philadelphia: E. Barrington and G.D. Haswell, 1850. http://resource.nlm.nih.gov/101505669
24. Guo J, Sun Z, Tang H, Jia X, Wang S, Yan X, Ye G, Wu G (2016) Hybrid optimization algorithm of particle swarm optimization and cuckoo search for preventive maintenance period optimization. Discret Dyn Nat Soc. https://doi.org/10.1155/2016/1516271
25. Hardavella J et al (2017) Top tips to deal with challenging situations: doctor–patient interactions. Breathe 13(2):129–135
26. Hong W et al (2011) SVR with Hybrid chaotic immune algorithm for seasonal load demand forecasting. Energies 4:960–977
27. Huang E et al (2003) Gene expression predictors of breast cancer outcomes. Lancet 361(9369):1590–1596
28. Kundra H, Sadawarti H (2015) Hybrid algorithm of cuckoo search and particle swarm optimization for natural terrain feature extraction. Res J Inf Technol 7(1):58–69
29. Lafourcade A et al (2018) Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the french e3n cohort. BMC Cancer 18(1):171
30. Meric F et al (2003) Positive surgical margins and ipsilateral breast tumor recurrence predict disease-specific survival after breast-conserving therapy. Cancer 97(4):926–933
31. Meystre S, Haug P (2006) Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 39(6):589–599
32. Partridge S et al (2005) MRI measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival. Am J Roentgenol 184(6):1774–1781
33. Sada Y et al (2016) Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care 54(2):e9-14
34. Sharma H, Rizvi M (2017) Prediction of heart disease using machine learning algorithms: A survey. Int J Recent Innov Trends Comput Commun 5(8):99–104
35. Shim H et al (2014) Breast cancer recurrence according to molecular subtype. Asian Pac J Cancer Prev 15(14):5539–44
36. Song W et al (2012) The risk factors influencing between the early and late recurrence in systemic recurrent breast cancer. J Breast Cancer 15(2):218–223
37. Stenkvist B et al (1982) Predicting breast cancer recurrence. Cancer 50(15):2884–2893
38. Tseng Y et al (2019) Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform 128:79–86
39. Vinitha S, Hao Y, Hwang K, Wang Lu, Wang Li (2019) Disease prediction by machine learning over big data from healthcare communities. Comput Sci Eng 8(1). https://doi.org/10.1109/ACCESS.2017.2694446
40. Young I, Luz S, Lone N (2019) A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 132(103971). https://doi.org/10.1016/j.ijmedinf.2019.103971
41. Yousefi M et al (2018) Organ-specific metastasis of breast cancer: molecular and cellular mechanisms underlying lung metastasis. Cell Oncol 41(2):123–140
42. Zhang Z, Hong W, Li J (2020) Electric load forecasting by hybrid self-recurrent support vector regression model with variational mode decomposition and improved cuckoo search algorithm. IEEE Access 8:14642–14658
43. Zhou M et al (2016) Discovery of potential prognostic long non-coding rna biomarkers for predicting the risk of tumor recurrence of breast cancer patients. Sci Rep 6(3):1038

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Amal Alzu'bi[1] • **Hassan Najadat**[1] • **Wesam Doulat**[2] • **Osama Al-Shari**[3] • **Leming Zhou**[4]

Hassan Najadat
najadat@just.edu.jo

Wesam Doulat
Wesam.dolat@gmail.com

Osama Al-Shari
omshari@just.edu.jo

Leming Zhou
Leming.Zhou@pitt.edu

[1] Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan

[2] Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

[3] Department of Internal Medicine-Clinical, Jordan University of Science and Technology, Irbid, Jordan

[4] Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA