Computational and Structural Biotechnology Journal xxx (2016) xxx–xxx

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning

Mohammad R. Mohebian [a], Hamid R. Marateb [a,b], Marjan Mansourian [c,*], Miguel Angel Mañanas [b], Fariborz Mokarian [d,e]

[a] Biomedical Engineering Department, Engineering Faculty, University of Isfahan, Hezar Jerib St., 81746-73441, Isfahan, Iran
[b] Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya, BarcelonaTech (UPC), C. Pau Gargallo, 5, 08028 Barcelona, Spain
[c] Department of Biostatistics and Epidemiology, School of Public Health, Isfahan University of Medical Sciences, Hezar Jerib St., 81745 Isfahan, Iran
[d] Cancer Prevention Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
[e] Department of Internal Medicine, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

Article history:
Received 30 August 2016
Received in revised form 24 November 2016
Accepted 26 November 2016
Available online xxxx

Keywords:
Breast cancer
Cancer recurrence
Computer-assisted diagnosis
Machine learning
Prognosis

ABSTRACT

Cancer is a collection of diseases that involves growing abnormal cells with the potential to invade or spread to the body. Breast cancer is the second leading cause of cancer death among women. A method for 5-year breast cancer recurrence prediction is presented in this manuscript. Clinicopathologic characteristics of 579 breast cancer patients (recurrence prevalence of 19.3%) were analyzed and discriminative features were selected using statistical feature selection methods. They were further refined by Particle Swarm Optimization (PSO) as the inputs of the classification system with ensemble learning (Bagged Decision Tree: BDT). The proper combination of selected categorical features and also the weight (importance) of the selected interval-measurement-scale features were identified by the PSO algorithm. The performance of HPBCR (hybrid predictor of breast cancer recurrence) was assessed using the holdout and 4-fold cross-validation. Three other classifiers namely as supported vector machines, DT, and multilayer perceptron neural network were used for comparison. The selected features were diagnosis age, tumor size, lymph node involvement ratio, number of involved axillary lymph nodes, progesterone receptor expression, having hormone therapy and type of surgery. The minimum sensitivity, specificity, precision and accuracy of HPBCR were 77%, 93%, 95% and 85%, respectively in the entire cross-validation folds and the hold-out test fold. HPBCR outperformed the other tested classifiers. It showed excellent agreement with the gold standard (i.e. the oncologist opinion after blood tumor marker and imaging tests, and tissue biopsy). This algorithm is thus a promising online tool for the prediction of breast cancer recurrence.

© 2016 Mohebian et al.. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computer-aided diagnosis (CAD) is using computers and software to interpret medical information. The purpose of CAD is to improve the diagnosis accuracy. In fact, CAD is used as a second opinion by the physicians to make the final diagnosis decision [1,2].

Nowadays, CAD is used in many different fields in medicine including, but not limited to, early detection of breast cancer [3], lung cancer diagnosis [4], arrhythmia detection [5] and dental and maxillofacial lesions diagnosis [6]. Several studies have been reported in the literature focusing on the application of CAD for cancer diagnosis and prognosis [7,8].

Cancer is a collection of diseases that involves growing abnormal cells with the potential to spread to other parts of the body [9]. There are over 200 types of cancer. The most common types of cancer in women are breast, colorectal, lung, and cervical [10].

Cancer is the leading cause of death worldwide, accounting for 8.2 million deaths in 2012 [11]. The most common causes of cancer death are: lung (1.59 million deaths), liver (745,000), stomach (723,000), colorectal (694,000), breast (521,000), and esophageal (400,000) [11]. More than 60% of world's total new annual cancer cases occur in Africa, and Central and South America [12]. The cancer incidence in Asia is also high [13]. It is expected that annual cancer cases

Abbreviations: CAD, computer-aided diagnosis; DT, decision tree; FH, family history of cancer; HPBCR, the proposed hybrid predictor of breast cancer recurrence; HRT, hormone therapy; I. Node, number of involved axillary lymph nodes; NR, lymph node involvement ratio; T. Node, number of dissected axillary lymph nodes; TS, tumor size; XRT, radiotherapy.

* Corresponding author.
E-mail addresses: van_mohebian@yahoo.com (M.R. Mohebian), h.marateb@eng.ui.ac.ir, hamid.reza.marateb@upc.edu (H.R. Marateb), j_mansourian@hlth.mui.ac.ir (M. Mansourian), miguel.angel.mananas@upc.edu (M.A. Mañanas), mokarian@med.mui.ac.ir (F. Mokarian).

http://dx.doi.org/10.1016/j.csbj.2016.11.004
2001-0370/© 2016 Mohebian et al.. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: Mohebian MR, et al, A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning, Comput Struct Biotechnol J (2016), http://dx.doi.org/10.1016/j.csbj.2016.11.004

UNCORRECTED PROOF

will rise from 14 million in 2012 to 22 million within the next 2 decades [11,12]. Worthwhile, among all types of cancer, breast cancer is the second leading cause of cancer death among women [14].

In Iran, cancer is the third cause of death after coronary heart disease, and accidents [15]. Breast cancer is the leading type of cancer in Iranian females, accounting for 24.6% of all cancers. In Iran, the average women age with breast cancer is 49.6 years [16]. Among all of the provinces in Iran, Isfahan Province is the biggest and most important area located at a desert border in the center of Iran [17]. The rate of cancer is increasing rapidly in Isfahan Province. Cancer control and comprehensive prevention plan is thus necessary [18].

Breast cancer, a complex heterogeneous disease, occurs with a set of different clinical symptoms. It is usually diagnosed using blood tests, MRI, mammography, and CT scan and biopsy. The pathology results from biopsy samples indicate whether the suspicious area is cancerous. Cancer patients, then, undergo a systematic treatment procedure dependent on the cancer stage and additional lab tests such as hormone receptor status [19].

Cancer staging (assigning an ordinal numbers I–IV) is in fact the process to determine the extent to which a cancer has been spread and is based on the following four characteristics: 1) the tumor size, 2) whether the cancer is invasive or non-invasive, 3) the spread of the tumor into the lymph nodes, and 4) the spread of the tumor into other parts of the body (i.e. metastasis). Stage I is an isolated cancer (tumor size ≤20 mm in greatest dimension) while stage IV is a metastasis cancer. Most cancer deaths are due to cancer that has spread from its primary site to other organs [12,20].

Breast cancer is usually treated with surgery, which may be followed by chemotherapy, radiation, and hormone therapies [21]. When the cancer patient is initially treated, the disease can recur at any time. However, most recurrences happen in the first 5 years after treatment [22]. The recurrence could be local (near the mastectomy scar), regional (spread to nearby lymph nodes) or metastatic (spread to other parts of the body not near the breast). Some of the most common sites of recurrence outside the breast are the lymph nodes, bones, liver, lungs, and brain [23].

An important issue is whether we can optimize the treatments to increase the therapeutic efficacy. In fact, 5-year recurrence-free survival is an important treatment quality measure. In principle, it is possible to predict 5-year cancer recurrence using clinicopathologic characteristics of cancer patients [24]. Such a prediction could be used by doctors to make proper treatment plan to considerably prolong patient life [25].

The prediction systems were used for cancer diagnosis in the literature [7]. However, there are few studies focusing on cancer prognosis (including recurrence or survival analysis). Since the focus of the current study is prediction of cancer recurrence, the literature review on cancer recurrence prediction models is provided. Meanwhile, the table of available methods was provided in the Supplementary material S1.

Zeng. suggested a mixture classification model containing a two-layer structure called mixture of rough set and support vector machine (SVM) for breast cancer prognosis with the average accuracy of 91% [26].

The Nottingham prognostic index (NPI) is a prognostic regression model proposed by Galea et al. [27] based on tumor size, histological grade, and lymph node status. The NPI calculation equation is as follows: tumor size (cm) × 0.2 + histological grade + lymph node point (negative nodes = 1; 1–3 positive nodes = 2; ≥4 positive nodes = 3). The patients were then classified into the low-risk (NPI point <3.4) and high-risk groups (NPI point ≥3.4) [27,28].

Kim et al. [28] used normalized mutual information index for feature selection and supported vector machines (SVM), Cox-proportional hazard regression model, and artificial neural network classifiers for classification in a sample size of 679 patients (the recurrence prevalence of 28.6%). The following features were used in their prognosis system: local invasion of tumor, number of tumors, number of metastatic lymph nodes, the histological grade, tumor size, estrogen receptor, and lymphovascular invasion and reached the sensitivity, specificity and area under the curve of 89%, 73% and 0.85, respectively for the best classifier (SVM). Although the statistical power of their system is acceptable (Power = 89% > 80%), the Type I error is beyond the acceptable range ($\alpha = 0.17 > 0.05$).

Ahmad et al. [29] used the SVM, decision tree, and multilayer perceptron artificial neural network classifiers with feature selection in a sample size of 547 patients (the recurrence prevalence of 21.4%). The predictors were age at diagnosis, menarche and menopause, tumor size, number of involved and dissected axially lymph nodes, grade and HER2. The best classifier (SVM) had sensitivity, specificity and accuracy of 96%, 91% and 94%, respectively. Despite the high performance of the algorithm proposed by the authors, cases with primary metastasis, a metastasis diagnosed at the time of registry, were not excluded from the analysis. In fact, having metastasis was one of the attributes primarily used by their proposed system. This necessary exclusion has been implemented in similar studies [28,30]. The reason for such exclusion is that the survival rate of patients with stage IV (metastasis) breast cancer within 5 years is about 10% to 15%. Thus, most of patients with primary metastasis experience cancer recurrence.

The discriminative prognosis predictors vary regionally in different studies [27–29]. Also, such prognosis tools have been recently used to optimize the treatment protocol. An example is including the cancer recurrence risk to breast cancer treatment guidelines by the American Society of Clinical Oncology (ASCO) and the National Comprehensive Cancer Network (NCCN). Thus, there is a need to design prognosis tools in regions that are different based on cancer epidemiology. Meanwhile, it is necessary to validate the cancer prognosis system using extensive diagnosis validation criteria. Therefore, the aim of this study was to *reliably* and *accurately* predict breast cancer recurrence using pathological and demographics features of the patients. In the proposed CAD system, HPBCR (hybrid predictor of breast cancer recurrence), a hybrid technique including statistical features selection, meta-heuristic population-based optimization and ensemble learning were used to predict breast cancer recurrence in the first 5 years after the diagnosis.

The rest of the paper is organized as follows: in the next section, information about the experimental protocol and the pattern recognition methods used in this study is presented. Section 3 provides the results of HPBCR and its comparison with the state-of-the-art. The discussion is provided in Section 4 and finally, the conclusions are summarized in Section 5.

2. Materials and Methods

2.1. The Cohort Dataset

In this study we used a 16-year registry cohort database (1998–2014) on 1085 women who diagnosed with breast cancer in Isfahan Sayed-o-Shohada cancer research center. History of clinical conditions and therapy of patients was continued until death or lost to follow up. Characteristics of variables and tumor have been recorded by interview and reported as pathology results. Time to recurrence was based on the physicians' opinion. The following information was extracted from each patient: age at diagnosis of breast cancer, lymph node involvement ratio (NR) defined as ratio of involved to dissected lymph nodes [31], age of menarche, number of pregnancy (No. Preg), primary tumor size (TS), cellular marker for proliferation (Ki67), number of involved (I. Node) and dissected (T. Node) axillary lymph nodes, number of chemotherapies (No. Chemo) and categorical covariates of family history of cancer (FH: negative/positive), having more than one tumor in the breast (multifocal: negative/positive), estrogen receptor expression (ER: negative (also known as absent)/positive), progesterone receptor expression (PR: negative (also known as absent)/positive), tumor protein 53 (p53: negative/positive), type of surgery (MRM: modified radical mastectomy, BCS: breast-conserving surgery, Mast: mastectomy), epidermal

UNCORRECTED PROOF

Please cite this article as: Mohebian MR, et al, A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning, Comput Struct Biotechnol J (2016), http://dx.doi.org/10.1016/j.csbj.2016.11.004

Download English Version:

https://daneshyari.com/en/article/8408337

Download Persian Version:

https://daneshyari.com/article/8408337

Daneshyari.com