Biomedical Signal Processing and Control 46 (2018) 86–93

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A quantitative performance study of two automatic methods for the diagnosis of ovarian cancer

Manuel A. Vázquez[a,b,1], Inés P. Mariño[c,d,*,1], Oleg Blyuss[e,d], Andy Ryan[d], Aleksandra Gentry-Maharaj[d], Jatinderpal Kalsi[d], Ranjit Manchanda[d,f], Ian Jacobs[d,g,h], Usha Menon[d,2], Alexey Zaikin[d,i,j,2]

[a] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés 28911, Madrid, Spain
[b] Gregorio Marañón Health Research Institute, Madrid 28009, Spain
[c] Department of Biology and Geology, Physics and Inorganic Chemistry, Universidad Rey Juan Carlos, Móstoles 28933, Madrid, Spain
[d] Department of Women's Cancer, Institute for Women's Health, University College London, London WC1E 6BT, United Kingdom
[e] Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London EC1M 6BQ, United Kingdom
[f] Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, United Kingdom
[g] Faculty of Medical and Human Sciences, University of Manchester, Manchester M13 9NT, United Kingdom
[h] Faculty of Medicine, UNSW Sydney, Sydney, NSW 2052, Australia
[i] Department of Mathematics, University College London, London WC1H 0AY, United Kingdom
[j] Department of Applied Mathematics, Lobachevsky State University of Nizhny Novgorod, Nizhniy Novgorod, Russia

ARTICLE INFO

Article history:
Received 31 March 2018
Received in revised form 30 May 2018
Accepted 3 July 2018

Keywords:
Ovarian cancer
Biomarkers
Deep learning
Recurrent neural networks
Markov chain
Monte Carlo
Gibbs sampling
Change-point detection
Bayesian estimation

ABSTRACT

We present a quantitative study of the performance of two automatic methods for the early detection of ovarian cancer that can exploit longitudinal measurements of multiple biomarkers. The study is carried out for a subset of the data collected in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). We use statistical analysis techniques, such as the area under the Receiver Operating Characteristic (ROC) curve, for evaluating the performance of two techniques that aim at the classification of subjects as either healthy or suffering from the disease using time-series of multiple biomarkers as inputs. The first method relies on a Bayesian hierarchical model that establishes connections within a set of clinically interpretable parameters. The second technique is a purely discriminative method that employs a recurrent neural network (RNN) for the binary classification of the inputs. For the available dataset, the performance of the two detection schemes is similar (the area under ROC curve is 0.98 for the combination of three biomarkers) and the Bayesian approach has the advantage that its outputs (parameters estimates and their uncertainty) can be further analysed by a clinical expert.

1. Introduction

Ovarian cancer remains the fifth most common cause of cancer-related deaths among women, with more than 150,000 annual deceases worldwide. Most cases occur in post-menopausal women (75%), with an incidence of 40 per 100,000 per year in women aged over 50. The early detection of this disease increases 5-year survival significantly, from 3% in Stage IV to 90% in Stage I [1]. Therefore, it is important to design efficient methods for early detection.

The screening and initial procedures for the detection of ovarian cancer are often carried out by testing serum biomarkers that are known to correlate with the appearance of tumours. In particular, the serum biomarker Canger Antigen 125 (CA125) is the most commonly used oncomarker in the screening of ovarian cancer [2–5]. However, other serum biomarkers have been reported to be associated with the development of ovarian cancer [6–8] and it has been recently suggested that they can be used in combination with CA125 [8–14]. The biomarker that has received more attention is the Human Epididymis Protein 4 (HE4), which has been used in the ROMA (Risk of Ovarian Malignancy Algorithm) to discriminate ovarian cancer from benign diseases [9,15] as well as in different panels for the purpose of early detection [7,10,11]. In a study within the Prostate Lung Colorectal and Ovarian (PLCO) cancer screening trial [16], HE4 was the second best marker after CA125, with a

* Corresponding author.
E-mail address: ines.perez@urjc.es (I.P. Mariño).
[1] These authors contributed equally to this work.
[2] These authors also contributed equally to this work.

https://doi.org/10.1016/j.bspc.2018.07.001
1746-8094/

sensitivity of 73% (95% confidence interval 0.60–0.86) compared to 86% (95% confidence interval 0.76–0.97) for CA125 [17,18]. Another serum biomarker, glycodelin, has also shown promising performance in the detection of ovarian cancer [12,19,20].

Recently, time series data from multiple biomarkers, including CA125, HE4 and glycodelin, have been jointly analysed to determine whether the level of these markers changed significantly and coherently at specific time instants [6], associating this fact with the development of tumours. The focus in [6] was placed on the detection of change-points for different biomarkers, by estimating the probability of coincidences as well as the probability of the change-point of a given biomarker appearing (and being detected) earlier than others. As a consequence, it was suggested that the combined detection of change-points in several biomarkers could be exploited for early diagnosis of ovarian cancer. In this paper we address the quantitative study of this automatic diagnostic technique using statistical analysis tools.

In particular, we study the trade-off between sensitivity (proportion of correctly detected positives) and specificity (proportion of correctly detected negatives) of a detection procedure that relies on the Bayesian change-point (BCP) model described in [6] which, in turn, is a version of the model proposed originally in [21] for the ROCA (Risk of Ovarian Cancer Algorithm) scheme. The quantitative analysis is carried out for a subset of the data collected in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) [22]. It involves time-series of CA125, HE4 and glycodelin for both healthy subjects (controls) and diagnosed patients (cases).

The decisions made by the BCP model involve estimating a number of parameters that admit a natural clinical interpretation. Although parsimony is always a desirable property to have in any model, accuracy (measured in terms of sensitivity and specificity) is here the ultimate goal. Hence, we also consider machine learning-based schemes which are often capable of modeling more complex mappings (between a set of measurements and the corresponding output) at the expense of some interpretability.

Deep learning (DL), and Recurrent Neural Networks (RNNs) in particular, have become important tools in classification tasks that involve the processing of ordered sequences of data [23]. Such methods have achieved state-of-the-art performance in applications such as handwriting [24], speech recognition [25] or image caption generation [26]. RNNS have also found many applications in the clinical field for tasks involving the classification of time series. In [27] a Long Short-Term Memory (LSTM) RNN is trained to classify diagnoses from pediatric intensive care unit (PICU) data. The same kind of data is fed to an RNN in [28] in order to predict mortality rates for patients in the intensive care unit. A Gate Recurrent Unit (GRU) is proposed in [29] for heart failure prediction. The authors of [30] use RNNs to assess the stress level of drivers from physiological signals coming from wearable sensors. In this work, we deploy a simple RNN for discriminating between women with ovarian cancer and healthy controls based on an ovarian cancer screening test that combines multiple biomarkers. The main challenge in applying DL in this context is the relatively small size of the dataset, which imposes some constraints on the kind of neural architectures that can be successfully trained without overfitting.

The ultimate goal in this paper is to carry out a comparison between these two different strategies (BCP and RNN) highlighting the advantages and disadvantages of both techniques. The study of both approaches, however, clearly shows that combining longitudinal time series of different biomarkers can improve the classification of pre-diagnosis samples regardless of the method.

The rest of this paper is organised as follows. Section 2 is devoted to the description of the dataset. Section 3 is devoted to a brief description of the Bayesian change-point method and the classification and statistical analysis carried out with it. In Section 4 the recurrent neural network technique is presented as well as the training procedure. The results obtained for both methods are presented and discussed in Section 5 and, finally, Section 6 is devoted to discussion and conclusions.

Table 1
Classification of cases, showing the range of ages and the average age over the corresponding women and samples.

Histology	Stages	Number of women	Range of ages	Average age
Serous cancers	I–II	9	[52.0–69.0]	61.3
	III–IV	18	[54.9–76.7]	66.6
Papillary	I–II	1	[68.1–69.2]	68.6
	III–IV	1	[55.2–57.2]	56.2
Endometrioid	I–II	2	[60.3–64.3]	62.7
	III–IV	1	[67.6–68.7]	68.1
Clear cell	I–II	2	[57.0–77.4]	67.2
	III–IV	0	0	0
Carcinosarcoma	I–II	0	0	0
	III–IV	3	[60.0–67.2]	63.7
Not specified cancers	I–II	2	[72.7–74.2]	73.5
	III–IV	5	[62.5–73.0]	67.8

2. Data

The two methods have been applied to a dataset from the multimodal arm [6] of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS, number ISRCTN22488978; NCT00058032) [22], where women underwent annual screening tests using the blood tumour marker CA125. Biomarkers HE4 and glycodelin assays were additionally performed on stored serial samples from a subset of women in the multimodal arm diagnosed with ovarian cancer and controls. The dataset included 179 controls (healthy women) and 44 cases (diagnosed women): 35 cases of invasive epithelial ovarian cancer (iEOC), 3 cases of fallopian tube cancer and 6 cases of peritoneal cancer. Out of these 44 cases, 16 are early stage (International Federation of Ginecology and Obstetrics, FIGO [31], stages I and II) and 28 are late stage (FIGO stages III and IV). In terms of histology, there are 27 serous cancers, 2 papillary, 3 endometrioid, 2 clear cell, 3 carcinosarcoma, and 7 not specified cancers. Each control has 4 to 5 serial samples available (177 controls with 5 samples and 2 controls with 4 samples) and each case has 2 to 5 serial samples available (24 cases with 5 samples, 10 cases with 3 samples and 10 cases with 2 samples). For healthy women, the range of age is 50.3–78.8 years and the average age over all women and samples is 63.6 years. On the other hand, the range of ages for cases is 52.0–77.4 years and the average age over all women and samples is 65.5 years. A detailed classification of the women with cancer is shown in Table 1, indicating the range of ages and the average age of the different subgroups.

All serum samples were assayed for CA125, glycodelin and HE4 using a proprietary multiplexed immunoassay based on Luminex technology which was developed and run by Becton Dickinson.

It should be noted here that all the biomarker measurements have been modified via a logarithmic transformation, as detailed in [12,21], in the form of $Y = \log(Z+4)$, where Z is the value of a particular marker.

Traditionally, single-biomarker time-series have been employed for the screening of ovarian cancer patients, particularly CA125 data. Recently, a few studies [6,12,32,33] have suggested that different biomarkers can be combined into multi-dimensional time-series and can lead to more accurate diagnosis. We explore this approach in the sequel.