This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

Using Contextual Learning to Improve Diagnostic Accuracy: Application in Breast Cancer Screening

Linqi Song, William Hsu, *Member, IEEE*, Jie Xu, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—Clinicians need to routinely make management decisions about patients who are at risk for a disease such as breast cancer. This paper presents a novel clinical decision support tool that is capable of helping physicians make diagnostic decisions. We apply this support system to improve the specificity of breast cancer screening and diagnosis. The system utilizes clinical context (e.g., demographics, medical history) to minimize the false positive rates while avoiding false negatives. An online contextual learning algorithm is used to update the diagnostic strategy presented to the physicians over time. We analytically evaluate the diagnostic performance loss of the proposed algorithm, in which the true patient distribution is not known and needs to be learned, as compared with the optimal strategy where all information is assumed known, and prove that the false positive rate of the proposed learning algorithm asymptotically converges to the optimum. In addition, our algorithm also has the important merit that it can provide individualized confidence estimates about the accuracy of the diagnosis recommendation. Moreover, the relevancy of contextual features is assessed, enabling the approach to identify specific contextual features that provide the most value of information in reducing diagnostic errors. Experiments were conducted using patient data collected at a large academic medical center. Our proposed approach outperforms the current clinical practice by 36% in terms of false positive rate given a 2% false negative rate.**

***Index Terms*—Computer-aided diagnosis system, online learning, contextual learning, multi-armed bandit, breast cancer.**

I. Introduction

Clinical decision support (CDS) tools help clinicians make detection and diagnostic decisions for complex diseases such as lung cancer [1], breast cancer [2][3], and diabetes [4]. There are a number of advantages to integrate CDS tools as part of the clinical workflow instead of solely relying on human intuition. First, the diagnostic accuracy of clinicians varies widely. A previous study has shown that false positive rates for breast cancer detection range from 2.6% to 15.9% among different radiologists, and younger and more recently trained radiologists have higher false-positive rates than experienced radiologists [5]; the deployment of CDS tools may reduce this variability. Second, although clinicians provide the correct diagnostic result in most cases, room for improvement exists in cases where discerning the difference between a benign or malignant mass is difficult [5]-[7]. CDS tools may provide better diagnostic recommendations in these cases by exploiting past knowledge of prior cases and their outcomes. Third, CDS tools help reduce fluctuations in diagnostic performance due to human factors (e.g., fatigue, distraction), offering consistent recommendations. Nevertheless, while these tools have been widely advocated to improve the diagnostic performance of clinicians, their adoption has remained limited given the following reasons: (1) the need to train current CDS tools using a fixed, predefined set of training cases; (2) the challenge of learning relevant features from high dimensional datasets; and (3) the inability to convey uncertainty that may be associated with a recommendation.

To address these challenges, we present an online algorithm for generating diagnostic recommendations to physicians by leveraging retrospective cases in the electronic health record (EHR) to reduce the false positive rate of diagnosis given a prescribed false negative rate. We demonstrate our approach in the domain of breast cancer screening and diagnosis, since breast cancer is a common cancer among women with an estimated 232,670 new cases among women in the United States in 2014 [8][9]. The proposed CDS tool is designed to aid physicians with making management decisions particularly in borderline cases. Radiological assessment of breast images are categorized using the BI-RADS (*Breast Imaging Report and Data System*) score. BI-RADS scores of 3 or 4 represent borderline cases associated with short interval followup or biopsy, respectively. Presently, many benign cases are being classified as BI-RADS 4A, which has raised the concern of overdiagnosis.

To improve the determination of whether a patient should be assigned as BI-RADS 3 or 4A, we explicitly consider the contextual information of the patient (also known as situational information) that affects diagnostic errors for breast cancer. The contextual information is captured as the current state of a patient, including demographics (age, disease history, etc.), the breast density (based on the BI-RADS breast density scale), the assessment history, whether the opposite breast has been diagnosed with a mass, and the imaging modality that was used to provide the imaging data. We hypothesize that the incorporation of contextual information will help provide more specific personalized diagnostic recommendations to patients.

The rest of the paper is organized as follows. Section II discusses related works. In Section III, we describe the system model and formulate the design problem. Section IV presents a systematic methodology for determining the optimal diagnostic recommendation strategy. Section V discusses practical issues related to the system: relevant context selection, prior information, and clinical regret. In Section VI, we present the experimental results and our findings. Section VII concludes

L. Song, J. Xu, and M. van der Schaar are with the Department of Electrical Engineering, UCLA, Los Angeles, CA 90095, USA. Email: songlinqi@ucla.edu, jiexu@ucla.edu, mihaela@ee.ucla.edu. This work was supported by the U.S. Air Force Office of Scientific Research under the DDDAS program.

W. Hsu is with the Department of Radiological Sciences, UCLA, Los Angeles, CA 90024, USA. Email: willhsu@mii.ucla.edu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

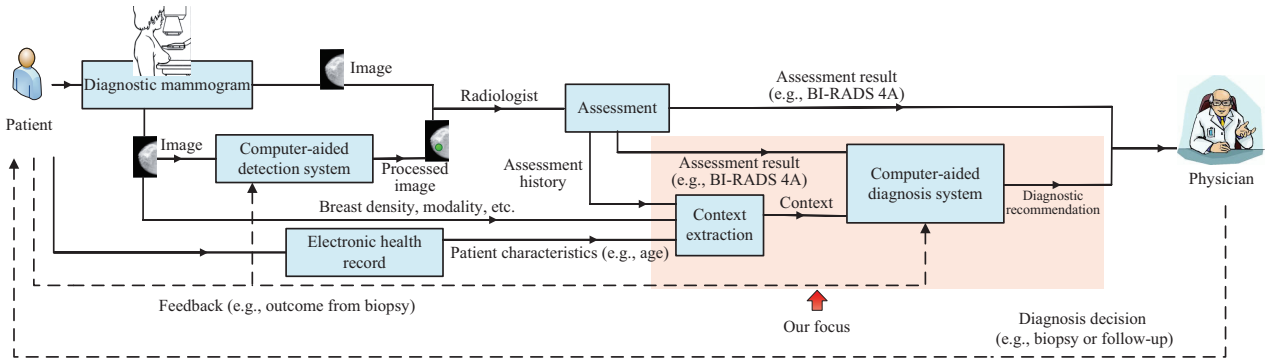

Fig. 1. The breast cancer diagnostic process overview.

TABLE I
A COMPARISON WITH PREVIOUSLY PUBLISHED WORKS.

	Employs BI-RADS	Adaptive strategy	Performance guarantee	Trade-off between FPR and FNR
[12]-[29]	No	No	No	No
[30]	Yes	No	No	No
Our work	Yes	Yes	Yes	Yes

the paper.

II. RELATED WORKS

A. Computer-Aided Detection and Diagnosis System for Breast Cancer

Various signal processing and machine learning techniques have been introduced to perform computer-aided detection and diagnosis. Early works have focused on image processing and classification techniques to extract features of the image and predict the outcome (i.e., whether benign or malignant) in the image [12]-[20]. A neural network-based algorithm [14] and a linear discriminant approach [2] are proposed to solve the diagnosis problem.

In breast screening, two types of CDS tools can be integrated in the diagnostic workflow, as depicted in Fig.1: (1) the computer-aided detection system, which helps the radiologist identify important features in the image that are abnormal [21]-[23]; and (2) the computer-aided diagnosis system, which helps physicians determine the diagnostic strategy for the patient (e.g., whether a patient should receive a biopsy or continue follow-up imaging) [24]-[30]. Our focus in this work is on the latter: we demonstrate how context derived from clinical (e.g., demographics, history) and imaging (e.g., breast density) sources can be used to provide diagnostic recommendations that reduce the number of biopsies performed while maintaining a low number of false negatives.

Clinically, management decisions that involve further interventions are indicated by a BI-RADS score of 0 (the imaging study cannot be interpreted and must be retaken), 4 (or 4A, 4B, and 4C, which represent different levels of suspicion), or 5 (high suspicion) [10][11]. A BI-RADS score of 3 means that the mass is likely benign with a short interval follow-up recommended. On the other hand, a BI-RADS 4 or 5 indicates that a biopsy is recommended. These decisions are made in the context of other clinical variables (e.g., if the mass is palpable). The difficulty lies in borderline cases (e.g., BI-RADS 3 or 4A) where indications are unclear whether a biopsy is truly necessary. For example, despite the cost and risk associated with biopsies, the positive predictive value for BI-RADS 4A is only 9% [34]. More efficient and accurate approaches are needed to reduce unnecessary biopsies [35]. In [30], a neural-fuzzy approach has been proposed, but these rule-based algorithms cannot be easily updated. Such approaches incur some performance loss because the underlying distribution of patient (outcome and context) is not known, and limited training data gives a limited estimation of the actual distribution, resulting in suboptimal diagnostic strategies. A partially observable Markov decision process (POMDP) has been used in [36] to solve a screening related question. However, the algorithm does not learn unknown distributions nor does it provide performance guarantees. The main components of our approach are illustrated in Fig. 1. Compared with existing works [2][3][30], our proposed framework employs an online learning approach, which continuously learns and adaptively updates the diagnostic strategy over time, eventually achieving the optimal strategy. Moreover, the proposed learning algorithm quickly (and provably) learns this optimal strategy. Learning in our framework is enabled by modeling each patient as characterized by his/her context and using the context to determine the similarity to the information gathered from other patients. Knowledge from diagnosis of former patients can only be transferred to the present/future patient by recognizing and exploiting similarities. Using contexts and their similarities, our approach is able to make diagnostic recommendations that are personalized to the patient. A comparison of our framework against existing frameworks is shown in Table I.

B. Contextual Multi-Armed Bandit

Our diagnostic recommendation algorithm is based on the contextual multi-armed bandit (MAB) framework [38][39][40] and incorporates the following innovations. First, prior information is considered, allowing the system to learn directly from prior information or other learners. Second, in existing works [38][39][40], the estimated error of an action can be updated only after the action is selected, and the algorithm needs to explore patients (by recommending different diagnostic actions to different patients under the same context) in order to get information about every action. However, due to ethical reasons, we cannot perform this type of exploration. In our algorithm, the diagnostic error of any action can be updated each time, and our algorithm does not need to explore patients in order to learn. Third, our algorithm considers minimizing the false positive rate, given a false negative rate constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

TABLE II
A COMPARISON WITH EXISTING MAB ALGORITHMS

	Explores patients	Considers prior information	Considers relevant context	Optimizes under constraint
Prior works [38]-[40]	Yes	No	No	No
Our approach	No	Yes	Yes	Yes

Existing works do not consider such a constraint. This is the first work using MABs for CDS and required several key innovations as compared to the conventional MAB works. A summary that compares the proposed learning approach with existing MAB works is shown in Table II.

III. SYSTEM MODEL

A. Computer-Aided Breast Cancer Diagnosis System

We consider a computer-aided breast cancer diagnosis system (CABCDS) as shown in Fig. 2. The system contains two modules: context extraction and computer-aided diagnosis. We consider a sequence of patients numbered $t = 1, 2, ...$ arrive with a borderline test result. Context extraction module aggregates information x_t from the EHR about a patient t, having a distribution of $f(x_t)$. Then, the computer-aided diagnosis module generates a diagnostic recommendation $\pi_t \in \{0,1\}$ to the physician, where 0 represents a 6-month imaging follow-up and 1 represents a biopsy. Here, we consider a binary decision, but the approach can easily be extended to incorporate additional choices.

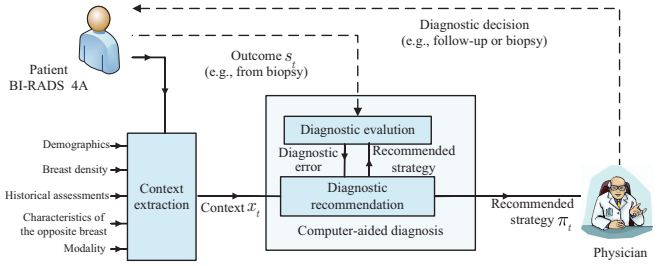

Fig. 2. Computer-aided breast cancer diagnosis system model.

B. Context Extraction Module

To better assist physicians, the CABCDS system considers a diverse set of contextual information to make sufficiently accurate recommendations. As shown in Table III, the following types of contextual features are considered: patient demographics (e.g., age, race), breast density, assessment history, whether the opposite breast has a high BI-RADS score previously (e.g., achieve BI-RADS 3), and imaging modality (e.g., mammogram, ultrasound). Regardless of whether the context variable is discrete or continuous, each context is modeled as discrete values that have numerical values between 0 and 1[1]. In this way, each context is a vector of features, or a point in the context space $\mathcal{X} = [0, 1]^{d_X}$, where d_X is the number of features. For example, a context can be represented as $x = (0.7, 0.2, 0, 1, 0)$, where $x_1 = 0.7$ represents a 70-year old patient with a scattered fibroglandular breast density ($x_2 = 0.2$), an assessment of only BI-RADS 1 or 2 ($x_3 = 0$) in the preceding screening study, an assessment of BI-RADS 3 for the opposite breast ($x_4 = 1$), and mammography as the imaging modality used ($x_5 = 0$).

[1]Note that the modeling of discrete contexts in different categories can also be considered as several learners, each corresponding to one category.

TABLE III
TYPES OF CONTEXTS AND DESCRIPTIONS

Context	Description
Demographics	The characteristics of a patient, age, race, disease history, family medical history, etc.
Breast Density	**Group 1**: The breast is almost entirely fat (fibrous and glandular tissue <25%). **Group 2**: There are scattered fibroglandular densities (fibrous and glandular tissue 25% to 50%). **Group 3**: The breast tissue is heterogeneously dense (fibrous and glandular tissue 50% to 75%). **Group 4**: The breast tissue is extremely dense (fibrous and glandular tissue > 75%).
Historical assessments	The information contained in previous imaging exam assessments (e.g., whether findings in BI-RADS 3 or higher appear in the past, or whether there is a significant change in the past year).
Characteristic of the opposite breast	The information of the opposite breast (e.g., whether findings in BI-RADS 3 or higher appear for the opposite breast).
Modality	The modality used for imaging: mammography (MG), ultrasound (US), magnetic resonance imaging (MRI) or computer radiography (CR).

C. Computer-aided Diagnosis Module

This module consists of recommendation generation and diagnostic evaluation steps. The recommendation generation step suggests a diagnostic strategy based on the contextual information and previous diagnostic evaluations. A diagnostic strategy is the approach for selecting an action, either to undergo a biopsy or to follow up, based on the observed contextual information. Given the context x_t of a patient t, $\pi_t(x_t)$ represents the action selected by the diagnostic strategy π_t. The strategy set is denoted by Π.

The diagnostic evaluation module collects outcomes of patients. The outcome of the patient t is $s_t(x_t)$, which is either 0 (representing benign) or 1 (representing malignant). If a patient undergoes a biopsy or returns for a short-term follow-up, the patient's outcome is revealed, where if the patient has been followed up for a certain time and the condition is stable, then the outcome is considered benign. We use $\sigma(x)$ to represent the probability of being malignant for a patient with context x. The evaluation of the diagnostic recommendation is through diagnostic errors. Two types of diagnostic errors are considered: false positive (e.g., if the outcome $s_t(x_t)$ is benign, and the recommended action is to undergo a biopsy) and false negative (e.g., if the outcome $s_t(x_t)$ is malignant, and the recommended action is a short-term follow-up).

D. Diagnostic Recommendation Problem

Based on the given CABCDS system, our design goal is to propose a recommendation algorithm that minimizes the false positive rate (FPR) given a tolerable false negative rate (FNR) η (e.g., $< 2\%$). The trade-off between false positive and false negative rates can be specified by a physician or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

by an institution. Therefore, the diagnostic recommendation problem is formally written as:

$$\begin{aligned} &\text{minimize} && \text{FPR} \\ &\text{subject to} && \text{FNR} \leq \eta \end{aligned} \tag{1}$$

IV. Diagnostic Recommendation Algorithm

The main idea of our diagnostic recommendation approach is to adaptively cluster patients based on related contexts and then learn the best action for each patient cluster.

A. Structure of the Optimal Strategy

In order to solve the diagnostic recommendation problem, we first analyze the structure of the optimal solution where all information (i.e., the distribution of context $f(x)$ and the probability of being malignant $\sigma(x)$) is known. We observe that the underlying probability $\sigma(x)$ varies for different contexts x, and hence, the solution is to recommend a biopsy for patients with a sufficiently high probability of having a malignancy, and to recommend a short interval follow-up for patients with a sufficiently low probability.

Proposition 1: The optimal strategy $\pi^*(x)$ for the diagnostic recommendation problem in eq. (1) is a threshold strategy: there exists a threshold σ_η, such that the optimal strategy satisfies $\pi^*(x) = 1$, if $\sigma(x) \geq \sigma_\eta$, and $\pi^*(x) = 0$, otherwise.

The intuition of this proposition is as follows: performing a biopsy when the probability of being malignant is low induces a high false positive rate. Conversely, performing a short-interval follow up when the probability of being malignant is high induces a high false negative rate.

Proof: See Appendix A. ■

Note that the context distribution $f(x)$ and outcome distribution $\sigma(x)$ are not known in practice. As such, the algorithm needs to learn the distribution of contexts and outcomes.

B. Description of the Proposed Learning Algorithm

Section III emphasized the need to uniquely characterize patients using contexts. However, no two patients are exactly the same. Knowledge from diagnosis and treatment of former patients can only be transferred to the present/future patient by recognizing and exploiting similarities among patients. Our proposed algorithm uses patients with similar contexts to accumulate information about a new patient and make recommendations based on the accumulated knowledge. As knowledge about more patients within a cluster becomes available, our algorithm adaptively shrinks the cluster size, thereby allowing more specific recommendations to be made.

The proposed algorithm maintains a set of disjoint clusters (referred to as "active clusters") that cover the whole context space. For each active cluster C, the algorithm maintains an estimate of the probability of a patient in this cluster of having a malignant tumor $\bar{\sigma}_C$. Based on these estimates of all active clusters, the algorithm finds the optimal threshold σ_t determined in Proposition 1 and the corresponding recommendation strategy for each cluster such that the false positive rate is minimized provided that the false negative rate is below the given tolerance η. After the patient outcome is revealed, the algorithm updates the estimate of the probability of a patient in cluster C of being malignant $\bar{\sigma}_C$. If the number of patient cases M_C belonging to a certain cluster C exceeds a certain threshold, the algorithm splits the cluster into smaller clusters in order to make more specific recommendations without sacrificing accuracy. We denote by $\mathcal{A}$ the set of active clusters, by $\bar{\sigma}_\mathcal{A}$ the set of $\bar{\sigma}_C$ in all clusters $C \in \mathcal{A}$, and by $M_\mathcal{A}$ the set of M_C in all clusters $C \in \mathcal{A}$.

The diagnostic recommendation algorithm is formally presented in Table IV and depicted in Fig. 3. When a patient arrives, the system extracts the context of the patient. For example, the patient is 60 years old and has a dense breast (for illustration, we only consider the two features of the context). The algorithm then finds an active cluster C, which the patient belongs to. Based on former patient cases, the algorithm determines the optimal threshold σ_t, and recommends the strategy: to undergo a biopsy if the estimated probability of malignancy for cluster C exceeds this threshold σ_t, and to follow-up otherwise. To determine the optimal threshold σ_t, the algorithm finds the highest value of σ_t such that the false negative rate is below the tolerant level η. If sufficient number of patient cases exists, the context space refinement process is performed that splits the current cluster into smaller clusters. The context space is uniformly partitioned [2] by the algorithm on each dimension (feature) by 2^l, each cluster (not necessarily the active cluster) with size 2^{-l}. The refinement process is to split an active cluster with size 2^{-l} into 2^{d_X} smaller clusters with size $2^{-(l+1)}$[3]. As shown in Fig. 3, when $d_X = 2$, a cluster with size $1/2$ is split into 4 clusters with size $1/4$. For example, consider a patient cluster in which patients are aged from 60 to 79 with breast density of Group 1 or 2. As new patients are added that fit this cluster, a more accurate estimation of the probability of malignancy can be characterized for patients in this cluster. When the patient cases are sufficiently many, the cluster is split into finer clusters in order to make more specific diagnostic recommendations: patients aged 60 to 69 with breast density Group 1, patients aged 60 to 69 with breast density Group 2, patients aged 70 to 79 with breast density Group 1, and patients aged 70 to 79 with breast density Group 2. After the diagnostic decision of the patient is made and the outcome of the patient is revealed, the patient case counter M_C and the estimated probability of being malignant $\bar{\sigma}_C$ are updated accordingly.

C. Evaluation of Algorithm Performance

In this subsection, the performance of the proposed DRA algorithm is analyzed in terms of the learning regret, which is the expected false positive rate of our learning algorithm compared with the optimal strategy $\pi^*(x)$, assuming all information (i.e., the distribution of context $f(x)$ and the probability of patient outcome $\sigma(x)$) is known. In practice, the information is not known and needs to be learned. We consider two types of regrets: individual patient regret (IPR) and aggregate system regret (ASR). The IPR is defined as the performance difference

[2]Other types of partitioning methods can also be applied.

[3]The splitting is determined by a size-dependent function $p(2^{-l}) = 2^{pl}$, where an empirical parameter p depends on how fast the clusters are to be partitioned. A smaller p results in a faster partition process of the context space, and a larger p results in a slower partition process of the context space.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

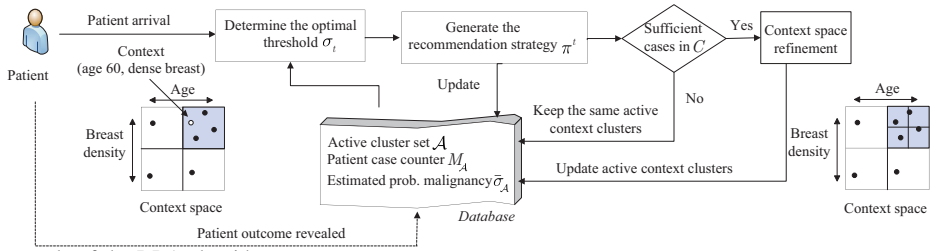

Fig. 3. An illustrative example of the DRA algorithm.

TABLE IV
DIAGNOSTIC RECOMMENDATION ALGORITHM (DRA)

diagnostic recommendation algorithm
Initialization: Set active cluster set $\mathcal{A} = \{\mathcal{X}\}$, patient case counter $M_C = 0$ for each active cluster C, estimation of probability of malignancy $\bar{\sigma}_C = 1$ for each active cluster C.

1: **for** $t = 1, 2, ...$ **do**
2: Context x_t arrives. Find the active cluster C, such that $x_t \in C$.
3: Determine the threshold σ_t:
$\sigma_t = threshold\_determination(\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}, t, \eta)$.
4: Output the recommendation strategy:
$\pi^t(C) = 0$, if $\bar{\sigma}_C < \sigma_t$, and $\pi^t(C) = 1$, otherwise.
5: The outcome s_t of the patient is revealed.
6: Update patient case counter for active cluster C: $M_C = M_C + 1$.
7: Update the estimation: $\bar{\sigma}_C = \frac{\#\ positive\ cases\ in\ C}{M_C}$.
8: **if** $M_C \geq p(2^{-l})$ **then** ▷ Context space refinement
9: Update the set of active clusters:
$(\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}) = context\_space\_refinement(\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}, C)$
10: **end if**
11: **end for**

$threshold\_determination(\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}, t, \eta)$
Input: active cluster set $\mathcal{A}$, patient case counter $M_{\mathcal{A}}$, estimation of probability of malignancy $\bar{\sigma}_{\mathcal{A}}$, the false negative rate tolerance η.
Output: the optimal threshold σ

1: Initialize the estimation of probability of malignancy $\sigma = 1$ and the false negative rate estimation $\bar{\mu}_1 = 1$.
2: **while** $\bar{\mu}_1 > \eta$ **do**
3: $\sigma = \sigma - t^{-1}$.
4: Calculate the false negative rate for threshold σ:
$\bar{\mu}_1 = \frac{\sum_{C:\bar{\sigma}_C \leq \sigma} M_C}{\max\{1, t-1\}}$.
5: **end while**
6: return σ.

$context\_space\_refinement(\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}, C)$
Input: active cluster set $\mathcal{A}$, patient case counter $M_{\mathcal{A}}$, estimation of probability of malignancy $\bar{\sigma}_{\mathcal{A}}$, the cluster C to be split.
Output: the updated $\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}$

1: Split the cluster C with size 2^{-l} into 2^{d_X} subclusters with size $2^{-(l+1)}$.
2: Remove cluster C from the active cluster set $\mathcal{A}$, and add the 2^{d_X} subclusters into the active cluster set $\mathcal{A}$.
3: Initialize the patient case counter and estimation of probability of malignancy for each subcluster C':
$M_{C'} = \#\ patient\ cases\ in\ C'$.
$\bar{\sigma}_{C'} = \frac{\#\ positive\ cases\ in\ C'}{M_{C'}}$;
4: return $\mathcal{A}, M_{\mathcal{A}}, \bar{\sigma}_{\mathcal{A}}$.

in terms of the false positive rate of strategy π_t and that of $\pi^*(x)$ for patient t, and the ASR is defined as the aggregated performance difference in terms of the false positive rate of strategy π_t and that of $\pi^*(x)$ for patients $1, 2, ...T$.

The performance of the proposed algorithm can be guaranteed in terms of the regret, as shown in Theorem 1.

Theorem 1: The ASR of the DRA algorithm up to time T can be bounded by $R(T) = O(T^{g(d_X)})$, and the IPR of the DRA algorithm for patient t can be bounded by $r(t) = O(t^{g(d_X)-1})$, where $0 < g(d_X) < 1$ is a parameter depending on the number of features d_X.

Proof: See Appendix B. ■

Note that, the IPR $O(t^{g(d_X)-1})$ goes to 0, as t goes to infinity, implying that the proposed DRA algorithm will converge to the optimal diagnostic performance. Accordingly, the following corollary is introduced.

Corollary 1: The performance of the proposed DRA algorithm converges to the optimal performance, in terms of the false positive rate.

In addition, from Theorem 1, the convergence speed is fast, at least in a sublinear[4] rate, as shown in Fig.5 and Fig.6 of the experimental results.

D. Algorithm Performance with Known Threshold

In previous sections, the algorithm is shown to adaptively learn the optimal threshold σ_η (probability of being malignant) over time, given a false negative rate constraint. However, in some clinical contexts, a fixed false negative rate may already exist (e.g., based on physician preference or clinical practice guidelines) [10] [11]. In this situation, the DRA algorithm degrades to a fixed threshold-based algorithm that does not need to learn the threshold value over time. Hence, step 5 of the DRA algorithm can be omitted, and the algorithm only needs to learn the distribution of patient outcomes $\sigma(x)$. We consider a weighted false positive and false negative error $c(\pi(x), s(x))$ in this setting:

$$c(\pi(x), s(x)) = \sigma_\eta * \text{false positive errors} + (1 - \sigma_\eta) * \text{false negative errors}. \quad (2)$$

The strategy for minimizing the expectation of the weighted error $c(\pi(x), s(x))$ is obtained by

$$\min_{\pi \in \Pi} Ec(\pi(x), s(x)). \quad (3)$$

The optimal strategy is denoted by $\pi^\dagger(x)$, which assumes all information is known, leading to the following proposition:

Proposition 2: The optimal strategy $\pi^\dagger(x)$ is equivalent to the optimal strategy $\pi^*(x)$ for the same σ_η.

Proof: Appendix C. ■

Intuitively, Proposition 2 shows that the optimal weighted error minimization strategy is equivalent to the optimal threshold-based strategy. Hence, we define regret as the difference in the weighted error, comparing our fixed threshold-based DRA algorithm to the optimal strategy $\pi^*(x)$. Formally, the ASR is defined as

$$R_\pi(T) = \sum_{t=1}^{T} [Ec(\pi^t(x_t), s(x_t)) - Ec(\pi^*(x_t), s(x_t))]. \quad (4)$$

We have the following theorem to bound this regret:

Theorem 2: The ASR of the fixed threshold-based DRA algorithm is bounded by $R(T) = O(T^{g(d_X)})$, and the IPR for patient t is bounded by $r(t) = O(t^{g(d_X)-1})$.

Proof: See Appendix D. ■

[4]A sublinear rate indicates that the expected performance loss is $O(1/t^\gamma)$ for patient t, where $0 < \gamma < 1$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

The regret in terms of weighted error of the fixed threshold-based DRA algorithm has the same sublinear order as the regret in terms of false positive rate of the DRA algorithm. This finding implies that the fixed threshold-based DRA algorithm will converge to the optimal diagnostic strategy, summarized by the following corollary.

Corollary 2: The performance of the fixed threshold-based DRA algorithm converges to the optimal performance, in terms of the expected weighted error in eq. (4).

In addition, based on Theorem 2, we can see that the convergence speed is fast (sublinear in the number of received patient cases).

V. Practical Considerations

In this section, we discuss some practical issues related to the system implementation and give appropriate approaches to address these issues.

A. Relevant Context Analysis

While large amounts of patient data are routinely captured in the EHR as part of clinical care, some information is inherently more relevant to assessing the probability of breast cancer than others. In an online learning setting, identifying which contextual information is more relevant to making a clinical recommendation based on retrospective data is important. For the d_X-dimensional context space $\mathcal{X}$, the probability of making diagnostic errors may be correlated with missing or noisy data. For example, the chance of making a diagnostic error may be low when results of a molecular assay is available, but the chance of making a diagnostic error may be high when only information about a patient's previous BI-RADS assessment is known.

For a d_X-dimensional context space, 2^{dx} DRA learning instances can be executed at the same time. At time t, the average false positive rates of all the learning instances are evaluated. We denote by instance 1 the learning instance using all d_X-dimensional contextual information. If for another learning instance i, the difference of false positive rate compared with instance 1 is below some level given by physicians, then the contextual information used by learning instance i is relevant. Hence, the system can select the more relevant context to make diagnostic recommendations, as shown in Fig. 4. From Theorem 1, when less contextual information is used, the convergence rate improves. This property is demonstrated in practice using actual data in Section VI-C.

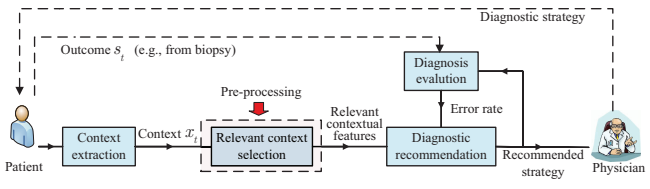

Fig. 4. The CABCDS system with relevant context.

B. Learning with Prior Information

Although the relevant context analysis can help identify significant predictors from the entire information space, the selection process can be challenging when little information is known about the underlying patient distribution, a problem known as "cold-start". One approach to solve this is to introduce prior contextual information, such as probabilistic statements that have been previously reported in other research studies, showing the relationship between contexts and the probability of cancer [37]. This prior information can be represented as an input parameter into the relevant context analysis module to derive an initial distribution.

Another source of prior information can be drawn from previously seen patient cases with known outcomes or reported in published literature. The prior statistical information includes the distribution of patients with cancer or the false negative rate and false positive rate obtained from other studies. The effect of using prior information is equivalent to a number of N training patient cases before running the system. In this case, the ASR can be bounded by $R(T) = O((T-N)^{g(d_X)})$. This shows that the performance of the system is greatly improved at the beginning of its operation since it can successfully capitalize on the prior knowledge.

C. Clinical Regret Analysis

In previous sections, we describe the algorithm for the CABCDS system and evaluated its performance in terms of learning regret (IPR or ASR). However, in practice, physicians may not always follow the system's recommendations due to differences in opinion between the experience of the physician and the recommended diagnostic strategy. In this section, we further evaluate the system performance by taking into account the actions of physicians and whether their actions are in agreement with the diagnostic recommendation. We call this analysis *clinical regret*.

We make the assumption that physicians have a certain but fixed probability of not following the recommended strategy when the system is first deployed. In this scenario, we denote the probability of not following the recommended strategy by ε. That is, when the diagnostic strategy recommended by the CABCDS system is a_t, the physician has a probability ε of selecting a strategy $\hat{a}_t \neq a_t$.

Theorem 3: Given a fixed probability ε of not following the recommended strategy, the clinical ASR up to time T can be bounded by

$$R(T) = O(T^{g(d_X)}) + O(\varepsilon T). \tag{5}$$

Proof: See Appendix E. ■

In this case, since the system performance converges to the optimal strategy, a constant probability of deviation will result in a linear clinical ASR in T. We now consider the scenario where the physicians have a decreasing probability of not following the recommended strategy given that confidence estimates for diagnostic recommendations increase as the system learns from additional cases. In this scenario, we denote the probability of not following the recommended strategy by $\varepsilon_t = \frac{1}{t^\beta}$ $(0 < \beta < 1)$. That is, when the diagnostic strategy recommended by the CABCDS system is π_t, the physician has a probability ε_t of selecting a strategy $\hat{\pi}_t \neq \pi_t$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

Theorem 4: Given a decreasing probability $\varepsilon_t = \frac{1}{t^\beta}$ of not following the recommended strategy, the clinical ASR up to time T can be bounded by

$$R(T) = O(T^{g(d_X)}) + O(T^{1-\beta}). \quad (6)$$

Proof: See Appendix E. ■

In this case, the clinical ASR has another sublinear term $O(T^{1-\beta})$. In fact, both the learning ASR of the DRA algorithm and the clinical ASR are sublinear in T, and hence converge to the optimal strategy, as shown in the following corollary.

Corollary 3: Given a decreasing probability $\varepsilon_t = \frac{1}{t^\beta}$ of not following the recommended strategy, the clinical performance in terms of false positive rate converges to the optimal performance.

VI. Experimental Results

In this section, the performance of the designed system is shown using our proposed algorithm. First, the breast cancer dataset used to evaluate the system performance is described. Then, our proposed online learning algorithm is evaluated and compared with other exiting algorithms. Finally, the impact of relevant contexts on the system performance in terms of diagnostic error rate and convergence rate is discussed.

A. Data Description

A de-identified dataset of 4,640 individuals who underwent screening and diagnostic mammograms at a large academic medical center is used. Patient outcome is derived from biopsy result, which is typically obtained for individuals with a BI-RADS score of 4 or 5. Our focus is on analyzing cases that are BI-RADS 4A; this category represents patients whose test results are less suspicious for cancer, raising the concern about unnecessary biopsies. We consider five contextual features, including:(1) patient age, (2) breast density, (3) assessment history (whether or not the immediately preceding exam shows a finding of BI-RADS 3 or above), (4) assessment results for the opposite breast (whether or not the immediately preceding exam shows a finding of BI-RADS 3 or above), and (5) the imaging modality used.

Characteristics of different BI-RADS categories are shown in Table V. The probability of being malignant increases from 9.91% to 78.61% as the BI-RADS category varies from 4A, 4B, to 4C. Prior to the introduction of BI-RADS 4A, 4B, 4C, all suspicious nodules were categorized as BI-RADS 4. The probability of being malignant of BI-RADS 4 is 26.12%, which is between those of BI-RADS 4A and 4B and near the total average probability of being malignant.

TABLE V
Description of Different Categories

BI-RADS	No. instances	Prob. of malignant
4	2282	26.12%
4A	1171	9.91%
4B	827	37.24%
4C	360	78.61%
Total	4640	28.08%

B. Performance Evaluation of the DRA algorithm

To perform the online adaptive learning, the data instances described previously are sequentially fed into the algorithm. Results are compared with the clinical approach and two other classical classifiers: the neural-fuzzy approach, and the linear discriminant analysis approach, which are defined as follows:

- **Clinical approach** [31]: Current clinical practice may be thought of as a threshold-based approach, which recommends a biopsy for all patients that fall in BI-RADS 4, 4A, 4B, 4C and above.
- **Neural-fuzzy approach** [14][30]: The neural-fuzzy approach models the diagnosis system as a three-layered neural network. The first layer represents input variables with various patient features; the hidden layer represents the fuzzy rules for diagnostic decision based on the input variables; and the third layer represents the output diagnostic recommendations.
- **Linear discriminant analysis (LDA)** [2][32]: The LDA approach trains a classifier using features extracted from imaging tests and assessment report, and the trained classifier can be used to make diagnostic recommendations.

System performance using different algorithms is shown in Fig. 5, Fig. 6, and Table VI. The false negative rate is empirically given as 5% and 2%, respectively. We assume that the same prior information or training data is available for each algorithm. Results show the relationship between average false positive rate and the percentage of patient arrivals (patient cases). The false positive rate of the proposed DRA algorithm decreases over time. In order to achieve a lower false negative rate, the false positive rate and the accuracy need to be sacrificed. Table VI shows that the false positive rate of the clinical approach is 1 for BI-RADS 4A patients, since it simply recommends all BI-RADS 4A patients to undergo a biopsy. The LDA algorithm cannot satisfy the false negative rate constraint, since it tries to linearly cluster patients based on the contextual information. However, the structure of the contextual information may not be linear, and as a result, a big performance loss is incurred. The neural fuzzy approach results in a performance loss and does not converge to optimum, likely because the trained rule may not be optimal and cannot be adaptively updated in time. Our DRA approach can be updated over time, maintaining a balance between false negative and false positive rates. The DRA algorithm outperforms the clinical approach in terms of the false positive rate by 39% and 36% for $\eta = 5\%$ and $\eta = 2\%$, respectively, the LDA approach in terms of the false positive rate by 34% and 30% for $\eta = 5\%$ and $\eta = 2\%$, respectively, and the neural-fuzzy approach in terms of the false positive rate by 14% and 12% for $\eta = 5\%$ and $\eta = 2\%$, respectively.

C. Contextual Feature Selection

The impact of contextual feature selection is twofold: (1) different contextual feature selections affect the performance of the algorithm, and (2) analysis of the selected features provides us with interesting insights of the underlying domain problem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

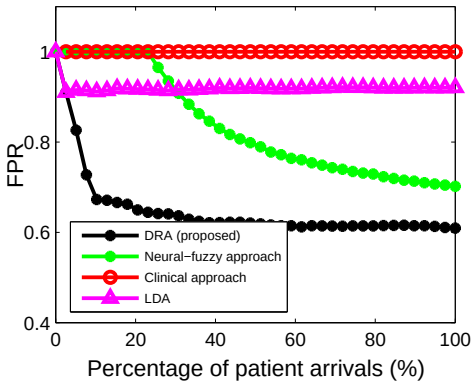

Fig. 5. Comparison of FPR for different algorithms, given tolerable FNR=5%.

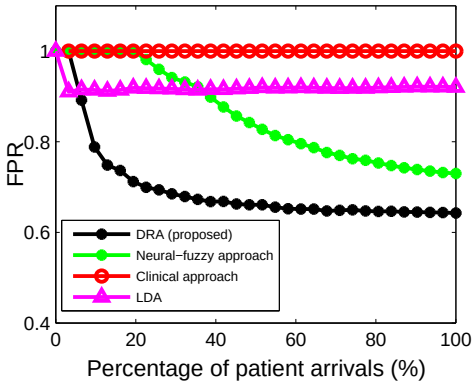

Fig. 6. Comparison of FPR for different algorithms, given tolerable FNR=2%.

Existing feature selection methods can be used as a preprocessing to the algorithm. We use a classical wrapper method, the branch and bound method, to select a subset of features [41] [42]. By comparing the scores (accuracy of classification) of different feature selections, this method selects the feature subset in a "backward elimination" manner: one starts with the set of all variables and progressively eliminates the least promising ones. We consider the scenario that 3 features are selected among all the 5 features. Results of using the branch and bound feature selection method and prediction accuracy of all possible subsets are shown in Table VII.

From the above simple example, we can see that different features influence the accuracy of recommendation. To illustrate the effect of context selection, we conduct the experiments using the proposed DRA algorithm for BI-RADS 4A patients by selecting different features and analyzing their relevance to answer the following three questions:

- Does using patient age information in addition to the BI-RADS results (breast density, assessment history of both breasts, modality, etc.) as contextual features improve the diagnostic accuracy?
- How do the breast density and the choice of imaging modality affect the diagnostic accuracy?
- Does the assessment history of patients provide valuable information to the diagnostic decisions?

The relevance of different contextual features to predict patient outcome is quantitatively described as the false positive rate and false negative rate.

(a) In Table VIII, we compare the results of using both age information and BI-RADS result information versus using only BI-RADS result information. Results show that no significant change in false positive rate is seen when the age information is considered. Although women with different ages have significantly different chances of having breast cancer [37], our results imply that the information about patient age plays a less important role in determining the diagnostic strategy than the BI-RADS test result information, such as breast density, assessment history, characteristic of opposite breast, and modality.

(b) In Table IX, we show the importance of considering breast density and modality in order to achieve a diagnostic recommendation by comparing the results of using both the breast density and modality, not using modality, and not using breast density or modality. Case 1 and Case 4 show that without the information about breast density and modality, the false positive rate increases by over 14% for both scenarios of 2% and 5% tolerable false negative rate. In addition, taking into account the information about breast density without knowing the modality can result in a significant increase in false positive rate, as shown by Case 1 and Case 3. In fact, no research has shown that breast density significantly implies the risk of cancer [10], but the breast density may cause lesions to be obscured in mammography [10][33]. Hence, using different modalities, such as mammography, ultrasound, and magnetic resonance imaging, can help reduce the diagnostic error when the patient has dense breasts.

(c) In Table X, we study the impact of assessment history of both breasts on determining the diagnostic strategy by comparing the results of using the assessment history of both breasts, using the same side breast without information about the opposite side breast, and not using the assessment history of any side breast. Results show that there is a 16% decrease in false positive rate when the tolerable false negative rate is low (i.e., 2%), and there is less than 7% variation in false positive rate when the tolerable false negative rate is high (i.e., 5%). Hence, the information about the assessment history of both breasts needs to be considered when a low false negative rate is suggested.

While these examples are illustrated using clinical features, the same approach can be extended to consider imaging features as well. The integration of imaging-derived features (e.g., texture, shape) is part of ongoing work.

TABLE VI
FPR AND FNR COMPARISON

Algorithms	$\eta = 5\%$			$\eta = 2\%$		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy
DRA (proposed)	61.0%	4.4%	0.45	64.2%	2.0%	0.42
Neural fuzzy	71.0%	4.9%	0.36	72.8%	1.9%	0.34
Clinical	100%	0	0.10	100%	0	0.10
LDA	92.1%	7.8%	0.16	92.1%	7.8%	0.16

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

TABLE VII
ACCURACIES OF DIFFERENT FEATURE SELECTIONS

	Subset of features									
	{1,2,3}	{1,2,4}	{1,2,5}	{1,3,4}	{1,3,5}	{1,4,5}	{2,3,4}	{2,3,5}	{2,4,5}	{3,4,5}*
Accuracy ($\eta = 5\%$)	0.34	0.33	0.38	0.32	0.37	0.39	0.33	0.38	0.37	0.41
Accuracy ($\eta = 2\%$)	0.28	0.20	0.28	0.29	0.33	0.32	0.24	0.24	0.32	0.37

{3,4,5}*: This is the subset of features obtained by the branch and bound algorithm.

As discussed in Section V, the different feature selections also affect the convergence rate of our online learning algorithm. In Fig. 7 and 8, we show results of convergence rate for the above discussed Case 7, where the assessment history of both breasts is not considered, Case 2, where the age information is not considered, and Case 1 with all contextual information. We can see from Fig. 7 that for a high tolerable false negative rate (5%), the convergence rate of Case 2 with a low context dimension is higher than that of the Case 7 with a medium context dimension, as well as than that of Case 1 with a high context dimension. However, for the low tolerable false negative rate (2%), the convergence rate of Case 7 is very low. In fact, as we previously showed, Case 7 does not consider the relevant contextual information regarding the assessment history of both breasts. This results in poor performance in terms of both the learning speed as well as the false positive rate in the scenario of a low tolerable false negative rate (2%).

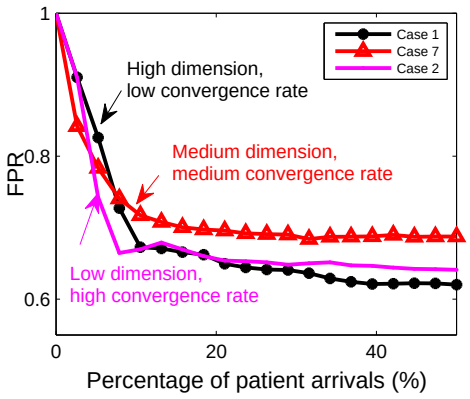

Fig. 7. Comparison of convergence rate for different context selection, tolerable FNR=5%.

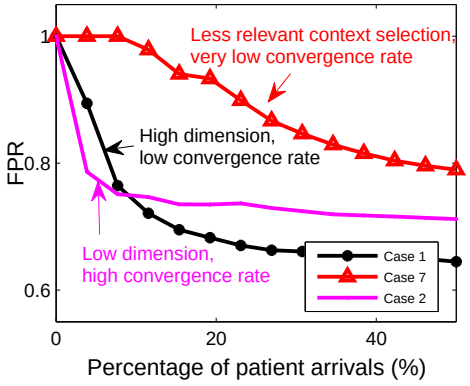

Fig. 8. Comparison of convergence rate for different context selection, tolerable FNR=2%.

D. Receiver Operating Characteristic of the System

To evaluate the system performance using different false negative rate tolerance, we simulate the receiver operating characteristic (ROC) of the system, as shown in Fig. 9. The goal of the ROC analysis is to show an overview of the system performance and the trade-off between the false positive rate and the false negative rate, according to which the physicians can determine the appropriate false negative rate tolerance level. As can be seen from Fig. 9, the false positive rate increases when the false negative rate decreases. In clinical practice, a balance between false positive and false negative rates need to be achieved: while we care more when a patient does not receive a timely treatment, the number of overdiagnosed cases must be minimized in order for CABCDS to be clinically accepted. We address this trade-off by minimizing the false positive rate given a user-defined false negative rate.

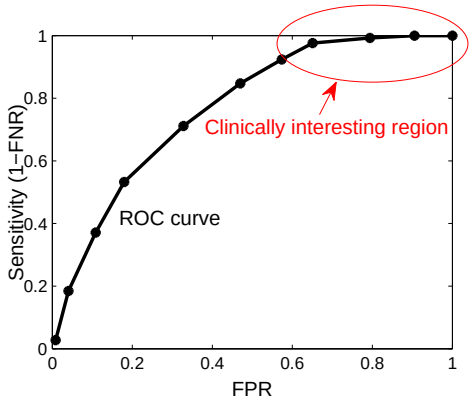

Fig. 9. Receiver operating characteristic of the proposed computer-aided diagnosis system.

VII. DISCUSSION AND FUTURE WORKS

This paper presents a novel design framework for a computer-aided breast cancer diagnosis system. The system incorporates contextual information and makes diagnostic recommendations to physicians, aiming to minimize the false positive rate of diagnosis, given a predefined false negative rate. The proposed algorithm is an online algorithm that allows the system to update the diagnosis strategy over time. We analytically show that the performance of our proposed algorithm converges to the optimal performance and quantify the rate of convergence.

The key contributions of this paper are:

- The process of breast cancer diagnosis is represented as a sequential decision making and online learning problem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

TABLE VIII
IMPACT OF AGE INFORMATION

Case	Contextual feature selctions					$\eta = 5\%$			$\eta = 2\%$		
	Age	Breast density	Assessment history	Opposite breast	Modality	FPR	FNR	Accuracy	FPR	FNR	Accuracy
1	X	X	X	X	X	61.0%	4.4%	0.45	64.2%	2.0%	0.42
2		X	X	X	X	63.1%	4.7%	0.43	67.2%	1.7%	0.40

TABLE IX
IMPACT OF BREAST DENSITY AND MODALITY

Case	Contextual feature selections					$\eta = 5\%$			$\eta = 2\%$		
	Age	Breast density	Assessment history	Opposite breast	Modality	FPR	FNR	Accuracy	FPR	FNR	Accuracy
1	X	X	X	X	X	61.0%	4.4%	0.45	64.2%	2.0%	0.42
3	X	X	X	X		72.1%	4.7%	0.35	81.0%	2.0%	0.27
4	X		X	X		74.8%	4.3%	0.32	79.6%	2.0%	0.29

TABLE X
IMPACT OF ASSESSMENT HISTORY OF BOTH BREASTS

Case	Contextual feature selections					$\eta = 5\%$			$\eta = 2\%$		
	Age	Breast density	Assessment history	Opposite breast	Modality	FPR	FNR	Accuracy	FPR	FNR	Accuracy
1	X	X	X	X	X	61.0%	4.4%	0.45	64.2%	2.0%	0.42
5	X	X	X		X	65.6%	4.7%	0.40	77.7%	1.8%	0.30
6	X	X		X	X	64.0%	4.8%	0.42	73.2%	2.0%	0.34
7	X	X			X	67.9%	4.4%	0.38	79.9%	1.9%	0.28

The Diagnostic Recommendation Algorithm (DRA) is formulated to make diagnostic recommendations over time while quickly converging on the optimal strategy. The algorithm exploits the dynamic nature of patient data (e.g., the context space grows as more patients are seen) to learn from and minimize the false positive rate of diagnosis given a false negative rate (e.g., $< 2\%$).

- Two types of "regret" (learning regret and clinical regret) are employed to evaluate the performance of the CDS tool. The regret associated with the proposed DRA algorithm is analytically quantified, showing that the false positive rate asymptotically converges to the optimal strategy and that the convergence rate is fast (i.e., sublinear).
- Selection of relevant contexts is performed in relation to minimizing diagnostic errors by identifying what knowledge or information is most influential in determining the correct diagnostic action. This information is provided to the physician who can decide what information to exploit so as to make efficient and effective diagnostic decisions.
- The proposed algorithm's performance is measured through experiments that incorporate clinical, imaging, and pathology data on 4,640 patients who underwent a diagnostic mammogram at our institution. Results show that an improvement in specificity can be achieved by exploiting the contextual information associated with the patient for the breast cancer diagnosis. Specifically, the proposed algorithm outperforms the current clinical approach by 36% in terms of the false positive rate given a 2% false negative rate.

One future work is to continue evaluating the presented framework and explore its implementation in the clinic. Understanding the utility and impact of the proposed approach in the current practice of breast cancer diagnosis requires further study. Nevertheless, the initial experimental results demonstrate that our online contextual learning algorithm is efficient, yet general and thus, it can potentially be used for diagnosis assist in other disease domains. Each of these domains has its own unique set of contextual information and desired patient outcomes. For example, in lung cancer screening patients, results of the low-dose computed tomography study (e.g., characterization of the nodule), pulmonary function tests, smoking and medical history, and environmental exposures are potential contexts. The contextual online learning algorithm can be adapted to handle such scenarios, helping physicians leverage available clinical big data to inform clinical decisions in each of these respective disease domains.

APPENDIX A
PROOF OF PROPOSITION 1

First, the recommended strategy is consistent: $\pi(x') \leq \pi(x'')$ if $\sigma(x') \leq \sigma(x'')$. The optimal solution will be among the threshold-based strategies: $\pi_\sigma(x) = 1$, if $\sigma(x) \geq \sigma$, and $\pi_\sigma(x) = 0$, otherwise. Second, the monotone property is satisfied: $E_x\mu_0(\pi_\sigma(x), s(x)) \leq E_x\mu_0(\pi_{\sigma'}(x), s(x))$ and $E_x\mu_1(\pi_\sigma(x), s(x)) \geq E_x\mu_1(\pi_{\sigma'}(x), s(x))$, when $\sigma \geq \sigma'$. Here we denote by μ_0 and μ_1 the false negative rate and the false positive rate. This implies that when a higher threshold is chosen, actions for some contexts will change from undergoing a biopsy to follow-up. Therefore, the false negative rate is reduced and the false positive rate is increased. Hence, a threshold σ_η exists, such that for any threshold-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

based strategy $\pi_\sigma(x)$ with $\sigma > \sigma_\eta$, the following property holds: $E_x\mu_1(\pi_\sigma(x), s(x)) > \eta$, and $E_x\mu_1(\pi_{\sigma_\eta}(x), s(x)) \leq \eta$. Obviously, the optimal solution is $\pi_{\sigma_\eta}(x)$, and we write the optimal solution as $\pi^*(x)$ for short.

APPENDIX B
PROOF OF THEOREM 1

We first describe the intuition of the proof, and then give the proof. The intuition is to cluster the contexts into small clusters over time. Within each context cluster, if $\sigma(x)$ within the context cluster has a gap from σ_η, and the estimation of $\sigma(x)$ is accurate enough, then the strategy in these context clusters are the same as the optimal strategy. The probability that the estimation of $\sigma(x)$ has a large deviation from the true value tends to 0. For the context clusters with $\sigma(x)$ close to σ_η, the strategies selected by the algorithm may not be the same as the optimal strategy, however, the probability of context arrivals in these clusters will tend to 0, since $\Pr\{x : \sigma(x) = \sigma_\eta\} = 0$. The strategy of the learning algorithm tends to the optimal strategy except some context clusters whose arrival probability tends to 0.

Formally, we define Lipschitz conditions for the theorem.

(1) **Patient outcome distribution**: there exists a Lipschitz constant $L_1 > 0$, such that for all $x, x' \in \mathcal{X}$, we have $|\sigma(x) - \sigma(x')| \leq L_1||x - x'||$.

(2) **Context distribution**: there exists a Lipschitz constant $L_2 > 0$, such that for all $x, x' \in \mathcal{X}$, we have $|f(x) - f(x')| \leq L_2||x - x'||$.

We consider the IPR r_t at some sufficiently large patient number t. Then we can see that the IPR can be decomposed into two terms $r_t = r_{t,1} + r_{t,2}$, where $r_{t,1}$ is the regret caused by clusters that have a $\sigma(x)$ near the threshold σ_η, and $r_{t,2}$ is the regret caused by clusters that have a $\sigma(x)$ far from the threshold σ_η, but have a wrong estimation of $f(x)$. We denote by $\sigma_{\min}(C) = \min_{x\in C}\sigma(x)$ the minimum probability of being malignant for cluster C, and denote by $\sigma_{\max}(C) = \max_{x\in C}\sigma(x)$ the maximum probability of being malignant for cluster C. We define three types of clusters:

Type I cluster: the cluster for patient t that has a $\sigma_{\max}(C)$ smaller than or equal to the threshold minus a small value, i.e., $\{C : C \in \mathcal{C}^t, \sigma_{\max}(C) \leq \sigma_\eta - bt^{-\alpha}\}$, where $b > 0$, $0 < \alpha < 1$ are parameters.

Type II cluster: the cluster for patient t that has a $\sigma_{\min}(C)$ greater than or equal the threshold plus a small value, i.e., $\{C : C \in \mathcal{C}^t, \sigma_{\min}(C) \geq \sigma_\eta + bt^{-\alpha}\}$.

Type III cluster: the remaining clusters that have a $\sigma(x)$ near σ_η, i.e., $\{C : C \in \mathcal{C}^t, \sigma_{\min}(C) - bt^{-\alpha} < \sigma_\eta < \sigma_{\max}(C) + bt^{-\alpha}\}$.

Due to Berstein's inequality, we have that the estimation for the context arrival at a cluster C has the following property:

$$\Pr\{|\frac{M_C}{t} - f(C)| > b_1 t^{1-\alpha}\} \leq b_{11}e^{-t^\alpha},$$

where b_1 and b_{11} are positive constants. And the realized $\bar{\sigma}_C$ in a cluster C has the following property:

$$\Pr\{\bar{\sigma}_C > \sigma_{\max}(C) + b_2 t^{1-\alpha} \text{ or } \bar{\sigma}_C < \sigma_{\min}(C) - b_2 t^{1-\alpha}\} \leq b_{22}e^{-t^\alpha},$$

where b_2 and b_{22} are positive constants. We define the normal state as the event that the estimations of $f(x)$ and $\sigma(x)$ are accurate enough. The set of normal states are denoted by $\mathcal{N}_C = \{|\frac{M_C}{t} - f(C)| \leq b_1 t^{1-\alpha}, \sigma_{\min}(C) - b_2 t^{1-\alpha} \leq \bar{\sigma}_C \leq \sigma_{\max}(C) + b_2 t^{1-\alpha}\}$. And we denote the set of abnormal state by $\bar{\mathcal{N}}_C$, which is the complementary set of $\mathcal{N}_C$. The probability of an abnormal state happens for one of the active cluster is bounded by $\sum_{C\in\mathcal{C}^t}\Pr\{\bar{\mathcal{N}}_C\}$.

Hence, we can see that the regret for $r_{t,1}$ is caused by Type III clusters, and can be bounded by

$$r_{t,1} \leq \Pr\{x : x \in C, \sigma_{\min}(C) - bt^{-\alpha} < \sigma_\eta < \sigma_{\max}(C) + bt^{-\alpha}\} \leq \frac{K2^{(d-1)l}}{2^{dl}} \leq K2^{-l},$$

where K is a constant, and the inequality is due to the covering property that the $d-1$ dimensional surface $\sigma(x) = \sigma_\eta$ and the Lipschitz condition that $\sigma_{\max}(C) - \sigma_{\min}(C) \leq L_1 2^{-l}$.

The regret $r_{t,2}$ can be bounded by the probability that an abnormal state $\sum_{C\in\mathcal{C}^t}\Pr\{\bar{\mathcal{N}}_C\}$ occurs. Hence, for patient t, the IPR can be bounded by

$$r_t = r_{t,1} + r_{t,2} \leq \sum_{C\in\mathcal{C}^t} K2^{-l} + b_{11}e^{-t^\alpha} + b_{22}e^{-t^\alpha} \leq O(t^{-1+g(d_X)}),$$

where $g(d_X) = \frac{d_X+1/2+\sqrt{9+8d_X}/2}{d_X+3/2+\sqrt{9+8d_X}/2}$, and the worst case context arrival and partition is considered as in Appendix D. Hence, we obtain the ASR up to patient T:

$$R(T) = \sum_{t=1}^{T} r_t \leq O(T^{g(d_X)}).$$

APPENDIX C
PROOF OF PROPOSITION 2

In order to show the equivalence of the two optimal strategies, we consider the weighted error of choosing different actions for context x. If the action $\pi(x) = 1$ is chosen, then

$$Ec(\pi(x) = 1, s(x)) = (1 - \sigma_\eta)\sigma(x). \tag{7}$$

If the action $\pi(x) = 0$ is chosen, then

$$Ec(\pi(x) = 0, s(x)) = \sigma_\eta(1 - \sigma(x)). \tag{8}$$

Hence, the optimal strategy $\pi^\dagger(x)$ satisfies:

$$\pi^\dagger(x) = \begin{cases} 1, \ if \ Ec(\pi(x) = 1, s(x)) \leq Ec(\pi(x) = 0, s(x)) \\ 0, \ otherwise \end{cases}. \tag{9}$$

By plugging (7) and (8) into (9), we have the optimal strategy $\pi^\dagger(x)$:

$$\pi^\dagger(x) = \begin{cases} 1, \ if \ \sigma(x) \geq \sigma_\eta \\ 0, \ otherwise \end{cases}. \tag{10}$$

Therefore, the proposition follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2414934, IEEE Journal of Biomedical and Health Informatics

APPENDIX D
PROOF OF THEOREM 2

To prove Theorem 2, we first introduce some important notions to characterize the properties of the cost. Let us define π_C^* as the best action corresponding to the context at the center of the subspace C. Let us also define $\mu_{x,\pi}$ as the expected weighted error, $\bar{\mu}_{C,\pi} = \max_{x\in C} \mu_{x,\pi}$, and $\mu_{C,\pi} = \min_{x\in C} \mu_{x,\pi}$. For a size 2^{-l} subspace C (referred to as "level l"), the *suboptimal action set* is defined as $\mathcal{L}_C(B) = \{\pi : \bar{\mu}_{C,\pi_C^*} - \mu_{C,\pi} > BLd_X^{\alpha/2}2^{-l\alpha}\}$. We add virtual exploration process into the algorithm: assume at least $2t^z \log(t)$ patients are diagnosed with error. We then can decompose the ASR into three terms: the regret caused by virtual exploration $R_e(T)$, the regret caused by suboptimal arm selection $R_s(T)$, and the regret caused by near optimal arm selection $R_n(T)$. We first introduce three lemmas to show useful properties of the DRA algorithm.

Lemma 1. The active cluster level l for patient t can be at most $(\log_2 t)/p + 1$.

Proof: According to the context space partition process, we have $\sum_{j=1}^{l} 2^{pj} \leq t$, where $l+1$ is the maximum level for patient t. Hence, the result follows.

Lemma 2. The regret caused by virtual exploration in one cluster up to patient t is bounded by $2t^z \log t$.

Proof: Since the virtual exploration number can be bounded by $t^z \log t$ for each action, the result follows.

Lemma 3. If $B = \frac{2}{Ld_X^{\alpha/2}2^{-\alpha}} + 2$, and $2\alpha p < z < 1$, then the regret caused by suboptimal action selection in one cluster up to patient t is bounded by $\frac{2\pi^2}{3}$.

Proof: Let $\mathcal{W}_C^t$ denote the event that the current phase is an exploitation phase in the context cluster C, and let $\mathcal{V}_C^t(\pi)$ be the event that the suboptimal action π is selected in at time t. Then, we have

$$\begin{aligned} R_{C,s}(T) &\leq \sum_{t=1}^{T} \sum_{\pi\in\mathcal{L}_C(B)} \Pr\{\mathcal{W}_C^t, \mathcal{V}_C^t(\pi)\} \\ &\leq \sum_{t=1}^{T} \sum_{\pi\in\mathcal{L}_C(B)} \Pr\{\bar{r}_{C,\pi} \geq \bar{\mu}_{C,\pi} + H_t, \mathcal{W}_C^t\} \\ &+ \Pr\{\bar{r}_{C,\pi_C^*} \leq \bar{\mu}_{C,\pi_C^*} - H_t, \mathcal{W}_C^t\} + \Pr\{\bar{r}_{C,\pi} \geq \bar{r}_{C,\pi_C^*}, \\ &\bar{r}_{C,\pi} < \bar{\mu}_{C,\pi} + H_t, \bar{r}_{C,\pi_C^*} > \bar{\mu}_{C,\pi_C^*} + H_t, \mathcal{W}_C^t\} \end{aligned} \quad , \tag{11}$$

where $H_t = t^{-z/2}$, $z \geq 2\alpha/p$. The third term on the right hand side of (11) is 0. Hence, we can bound the regret by

$$\begin{aligned} R_{C,s}(T) &\leq \sum_{t=1}^{T} \sum_{\pi\in\mathcal{L}_C(B)} \Pr\{\bar{r}_{C,\pi} \geq E[\bar{r}_{C,\pi}] + Ld_X^{\alpha/2}2^{-l\alpha}\} \\ &+ \Pr\{\bar{r}_{C,\pi_C^*} \leq E[\bar{r}_{C,\pi_C^*}] - Ld_X^{\alpha/2}2^{-l\alpha}\} \leq \sum_{t=1}^{T} 4t^{-2} \leq \frac{4\pi^2}{3} \end{aligned} \quad . \tag{12}$$

We can see that the highest level of subspaces is at most $1 + \log_{2^{p+d_X}} T$. Then the maximum number of subspaces is bounded by $2^{2d_X} T^{\frac{d_X}{d_X+p}}$.

Therefore, according to Lemmas 2, we can bound the exploration regret by

$$R_e(T) \leq 2^{2d_X+1} T^{\frac{d_X}{d_X+p}} T^z \log T. \tag{13}$$

Accord to Lemma 3, we can bound the suboptimal regret by

$$R_s(T) \leq \frac{2^{2d_X+1}\pi^2 T^{\frac{d_X}{d_X+p}}}{3}. \tag{14}$$

We can also bound the near optimal regret by

$$\begin{aligned} R_n(T) &\leq \sum_{l=0}^{1+\log_{2^{p+d_X}} T} BLd_X^{\alpha/2}2^{-l\alpha} \\ &\leq BLd_X^{\alpha/2}2^{2(d_X+p-\alpha)} T^{\frac{d_X+p-\alpha}{d_X+p}}. \end{aligned} \tag{15}$$

Therefore, the ASR follows by setting $z = 2\alpha/p$, and $p = \frac{3\alpha+\sqrt{9\alpha^2+8\alpha d_X}}{2}$. Since the IPR decreases as t increases, it can be bounded by $O(t^{g(d_X)-1})$.

APPENDIX E
PROOF OF THEOREM 3 AND 4

For Theorems 3 and 4, the clinical ASR can be calculated by another deviating regret term. For Theorem 3, this term can be bounded by $\sum_{t=1}^{T} \varepsilon = \varepsilon T$. For Theorem 4, this term can be bounded by $\sum_{t=1}^{T} \varepsilon_t \leq \frac{T^{1-\beta}}{1-\beta}$. Hence, Theorems 3 and 4 follow.

REFERENCES

[1] M. N. Gurcan, B. Sahiner, N. Petrick, H. P. Chan, E. A. Kazerooni, P. N. Cascade, et al., "Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system," *Medical Physics*, vol. 29, no. 11, pp. 2552-2558, 2002.

[2] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Transactions on Information Technology in Biomedicine,* vol. 13, no. 2, pp. 236-251, 2009.

[3] K. Ganesan, U. Acharya, C. K. Chua, L. C. Min, K. Abraham, and K. Ng, "Computer-aided breast cancer detection using mammograms: a review," *IEEE Reviews in Biomedical Engineering*, vol. 6, pp. 77-98, 2013.

[4] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14650-14659, 2011.

[5] J. G. Elmore, D. L. Miglioretti, L. M. Reisch, M. B. Barton, W. Kreuter, C. L. Christiansen, et al., "Screening mammograms by community radiologists: variability in false-positive rates," *Journal of the National Cancer Institute*, vol. 94, no. 18, pp. 1373-1380, 2002.

[6] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," *Biomedical informatics*, Springer, London, pp. 643-674, 2014.

[7] K. Kerlikowske, P. A. Carney, B. Geller, M. T. Mandelson, S. H. Taplin, K. Malvin, et al., "Performance of screening mammography among women with and without a first-degree relative with breast cancer," *Annals of Internal Medicine*, vol. 133, pp. 855–863, 2000.

[8] " Cancer Facts & Figures 2014," American Cancer Society, 2014.

[9] R. Seigel, D. Naishadham, and A. Jemal, "Cancer Statistics," American Cancer Society, 2013.

[10] "ACR BI-RADS breast imaging and reporting data system: breast imaging Atlas 5th Edition," American College of Radiology, 2013.

[11] L. Liberman and J. H. Menell, "Breast imaging reporting and data system (BI-RADS)," *Radiologic Clinics of North America*, vol. 40, no. 3, pp. 409-430, 2002.

[12] M. L. Giger, Z. Huo, M. A. Kupinski, and C. J. Vyborny, "Computer-aided diagnosis in mammography," *Handbook of medical imaging*, vol. 2, pp. 915-1004, 2000.

[13] D. Gur, J. H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. Hardesty, et al., "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *Journal of the National Cancer Institute*, vol. 96, no. 3, pp. 185-190, 2004.

[14] D. Tsai, H. Fujita, K. Horita, T. Endo, C. Kido, and T. Ishigaki, "Classification of breast tumors in mammograms using a neural network: utilization of selected features," In *Proc. IEEE International Joint Conference on Neural Networks*, pp. 967-970, 1993.

[15] F. Schnorrenberg, C. S. Pattichis, K. C. Kyriacou, and C. N. Schizas, "Computer-aided detection of breast cancer nuclei," *IEEE Transactions on Information Technology in Biomedicine*, vol. 1, no. 2, pp. 128-140, 1997.

[16] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, et al., "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 811-819, 1997.

[17] N. R. Mudigonda, R. M. Rangayyan, and J. L. Desautels, "Gradient and texture analysis for the classification of mammographic masses," *IEEE Transactions on Medical Imaging*, vol. 19, no. 10, pp. 1032-1043, 2000.

[18] Y.-H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, et al., "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," *Medical physics*, vol. 28, no. 4, pp. 455-461, 2001.

[19] C. M. Kocur, S. K. Rogers, L. R. Myers, T. Burns, M. Kabrisky, J. W. Hoffmeister, et al., "Using neural networks to select wavelet features for breast cancer diagnosis," *IEEE Engineering in Medicine and Biology Magazine*, vol. 15, no. 3, pp. 95-102, 1996.

[20] A. Urmaliya and J. Singhai, "Sequential minimal optimization for support vector machine with feature selection in breast cancer diagnosis," In *IEEE Second International Conference on Image Information Processing (ICIIP),* pp. 481-486, 2013.

[21] M. Sameti, R. K. Ward, J. Morgan-Parkes, and B. Palcic, "Image feature extraction in the last screening mammograms prior to detection of breast cancer," *IEEE Journal Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 46-52, 2009.

[22] M. Donelli, I. J. Craddock, D. Gibbins, and M. Sarafianou, "A three-dimensional time domain microwave imaging method for breast cancer detection based on an evolutionary algorithm," *Progress In Electromagnetics Research M*, vol. 18, pp. 179-195, 2011.

[23] P. S. Pawar and D. R. Patil, "Breast Cancer Detection Using Neural Network Models," In *IEEE International Conference on Communication Systems and Network Technologies (CSNT),* pp. 568-572, 2013.

[24] S. Timp, C. Varela, and N. Karssemeijer, "Computer-aided diagnosis with temporal analysis to improve radiologists' interpretation of mammographic mass lesions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 803-808, 2010.

[25] S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," In *IEEE First International Conference on Automation, Control, Energy and Systems (ACES),* pp. 1-4, 2014.

[26] S. Singh and K. Bovis, "An evaluation of contrast enhancement techniques for mammographic breast masses," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 1, pp. 109-119, 2005.

[27] K. Panetta, Y. Zhou, S. Agaian, and H. Jia, "Nonlinear unsharp masking for mammogram enhancement," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 6, pp. 918-928, 2011.

[28] A. N. Karahaliou, I. S. Boniatis, S. G. Skiadopoulos, F. N. Sakellaropoulos, N. S. Arikidis, E. A. Likaki, et al., "Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications, *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 731-738, 2008.

[29] G. Bozza, M. Brignone, and M. Pastorino, "Application of the no-sampling linear sampling method to breast cancer detection," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2525-2534, 2010.

[30] A. Keles, A. Keles, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5719-5726, 2011.

[31] W. Hsu, S. Han, C. Arnold, A. Bui, D. R. Enzmann, "RadQA: A data-driven approach to assessing the accuracy and utility of radiologic interpretations," *SIIM Annual Meeting*, 2014.

[32] K. Armstrong, E. A. Handorf, J. Chen, and M. N. B. Demeter, "Breast cancer risk prediction and mammography biopsy decisions: a model-based study," *American Journal of Preventive Medicine*, vol. 44, no. 1, pp. 15-22, 2013.

[33] S. Venkataraman, P. J. Slanetz, "Breast imaging: mammography and ultrasonography," *UptoDate*, 2014.

[34] C. Wiratkapun, W. Bunyapaiboonsri, B. Wibulpolprasert, and P. Lertsithichai, "Biopsy rate and positive predictive value for breast cancer in BI-RADS category 4 breast lesions," *Medical Journal of the Medical Association of Thailand*, vol. 93, no. 7, pp. 830, 2010.

[35] C. I. Flowers, C. O'Donoghue, D. Moore, A. Goss, D. Kim, J.-H. Kim, et al., "Reducing false-positive biopsies: a pilot study to reduce benign biopsy rates for BI-RADS 4A/B assessments through testing risk stratification and new thresholds for intervention," *Breast Cancer Research and Treatment*, vol. 139, no. 3, pp. 769-777, 2013.

[36] T. Ayer, O. Alagoz, and N. K. Stout, "OR forum-a POMDP approach to personalize mammography screening decisions," *Operations Research*, vol. 60, no. 5, pp. 1019-1034, 2012.

[37] S. W. Fletcher, "Risk prediction models for breast cancer screening," *UptoDate*, 2014.

[38] A. Slivkins, "Contextual bandits with similarity information," arXiv preprint arXiv:0907.3986, 2009.

[39] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1096-1103, 2007.

[40] T. Lu, D. Pal, and M Pal, "Contextual multi-armed bandits," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 485-492, 2010.

[41] R. Kohavi, and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273-324, 1997.

[42] P. M. Narendra, and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Tran. Computers*, vol. C-26, no. 9, pp. 917-922, 1977.