Predicting breast screening attendance using machine learning techniques

Baskaran, V. , Guergachi, A. , Bali, R. and Naguib, R.

Author post-print (accepted) deposited in CURVE January 2012

Original citation & hyperlink:

Baskaran, V. , Guergachi, A. , Bali, R. and Naguib, R. (2011) Predicting breast screening attendance using machine learning techniques. IEEE Transactions on Information Technology in Biomedicine, volume 15 (2): 251-259.

http://dx.doi.org/10.1109/TITB.2010.2103954

Predicting Breast Screening Attendance Using Machine Learning Techniques

Vikraman Baskaran, Aziz Guergachi, *Member, IEEE*, Rajeev K. Bali, *Senior Member, IEEE*, and Raouf N. G. Naguib, *Senior Member, IEEE*

***Abstract*—Machine learning-based prediction has been effectively applied for many healthcare applications. Predicting breast screening attendance using machine learning (prior to the actual mammogram) is a new field. This paper presents new predictor attributes for such an algorithm. It describes a new hybrid algorithm that relies on back-propagation and radial basis function-based neural networks for prediction. The algorithm has been developed in an open source-based environment. The algorithm was tested on a 13-year dataset (1995–2008). This paper compares the algorithm and validates its accuracy and efficiency with different platforms. Nearly 80% accuracy and 88% positive predictive value and sensitivity were recorded for the algorithm. The results were encouraging; 40–50% of negative predictive value and specificity warrant further work. Preliminary results were promising and provided ample amount of reasons for testing the algorithm on a larger scale.**

***Index Terms*—Breast screening, cancer, machine learning, neural networks, prediction, screening attendance.**

I. Introduction

Breast cancer is the most common cancer for women in North America [1]. In the U.K., over 40 000 women are being diagnosed with breast cancer each year [2], [3]. Mortality due to breast cancer is also one of the highest in the world [1], [4], and is the second highest of all cancers in the Canada [7]. Breast cancer should ideally be diagnosed at the earlier stages of its development to considerably reduce mortality. Possible treatments include removing or destroying the cancer cells to avoid the spread of the affected cells. Breast self-examination is an effective and noninvasive type of checking for any lumps in the breast tissue. Unfortunately, this greatly depends on the size of the lump, technique, and experience in carrying out a self-examination procedure by a woman [9]. An ultrasound test, examining breast tissue using sound waves, can be utilized to detect lumps but this is usually suited for women aged below 35 owing to the higher density of breast tissue [1]. Having a tissue biopsy via a fine needle aspiration or an excision is often used to examine the cells histopathologically and to diagnose if the growth, lump, is benign or cancerous. These investigations are mostly employed in treatments or post-treatment examination and as second rung diagnostic confirmation methods [10]. Performing a computed tomography or an MRI scan would result in a thorough examination of the breast tissue but these techniques are not favored due to reasons which include cost, needs preparation, noise, time, and images that may not be clear [10].

Mammography is a technique for detecting breast tissue lumps using a low dosage of X-ray. This technique can even detect a 3-mm-sized lump. The X-ray image of the breast tissue is captured and the image is thoroughly read by experienced radiologists and specialist mammogram readers [10]. Preliminary research suggests that women aged 50 and above are more susceptible to breast cancer; mammography is more suited to women in this age range due to the lower density of breast tissue [11]. Even though mammography has its critics—mainly due to its high rate of false positives and false negatives [13]—it has become the standard procedure for screening women by the NHS National Breast Screening Program in the U.K. [3], [15]. Mammography is the best and most viable tool for mass screening to detect cancer in the breast at an early stage [17]; however, the effectiveness of diagnosis through screening is directly dependent on the percentage of women attending the screening program [18]–[20]. The NHS Breast Screening Program, catering to the entire eligible women population, is funded by the Department of Health in the U.K. It covers 2.5 million women every year and detected nearly 16 500 cancers in the screened population for the year 2007–2008 [3]. Currently, the screening program routinely screens women between the ages 50 and 70.

Early breast cancer detection through screening is fundamental for increasing the efficacy of cancer treatment [11], [21]. Mammography has been accepted as the best and most economically viable tool for population screening [22]. Maximizing coverage for the target population is crucial for the success of such screening programs [11]. Currently, the breast cancer screening attendance rates are below expectations in many countries that have publicly funded healthcare programs [24]. This paper proposes a set of protocols to increase breast screening attendance for the U.K.'s NHS breast screening program. Based on this protocol, a new software prototype was created and tested. The prototype tests the prediction algorithm and shares the prediction results with multiple healthcare stakeholders for initiating opportunistic interventions on nonattendees. This prototype is a radical new idea that uses machine learning techniques for

Manuscript received April 11, 2010; revised October 12, 2010; accepted December 23, 2010. Date of publication; date of current version. This work was supported in part by the NHS Cancer Screening Programs, U.K.

V. Baskaran and A. Guergachi are with Ryerson University, TRSM, Toronto, ON M5B 2K3, Canada (e-mail: vikraman@ryerson.ca; a2guerga@ryerson.ca).

R. K. Bali is with the KARMAH Group, Health Design and Technology Institute (HDTI), Coventry University, Coventry University Technology Park, Coventry CV1 2TT, U.K. (e-mail: r.bali@coventry.ac.uk).

R. N. G. Naguib is with the Biomedical Computing and BIOCORE, HDTI, Coventry University, Coventry University Technology Park, Coventry CV1 2TT, U.K. (e-mail: r.naguib@coventry.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2010.2103954

predicting screening attendance and shares this knowledge by adopting the health informatics initiative of the NHS.

II. Challenge

The NHS Breast Screening Program Annual Review (2008) states that, out of invited women, only 74% attend the screening program [3]. This sizeable nonattendance could result in missed cancer detection for nearly 4 000 women (based on the cancer detection rate within screened women) [3]. This large percentage of nonattendance not only result in loss of life due to breast cancer but also result in loss of screening resources through costly imaging equipment laying idle, underutilization of specialist-imaging expertise, wasted screening slots, and so forth. Screening units are unable to arrange buffered attendees for the idle slots since the units do not know *a priori* which women will attend and which will not. In addition, there is a sizeable cost factor involved in sending repeat screening appointments letters to nonattending women.

Reasons for nonattendance may be largely attributed to disinterest in attending a mammography session, prior or current medical problems, and fear of X-rays [11], [24]. These reasons can be negated by proper education provided to women. Education has to be directed at explaining the advantages and importance of screening and assist in removing the sociocultural and personal barriers [25]. Other possible options include convenience in terms of time, place, and dates provided to women for encouraging their attendance.

In spite of the expedient measures provided to the women, nonattendance has been a grave concern for the NHS—National Screening Program. This scenario can be properly addressed if those women who may probably not attend a screening appointment can be identified in advance so that additional resources can be directed at interventions that can increase screening attendance.

A proposal enumerating the complete software solution is summarized at the end of Section IV. The National Screening Program has been constantly striving to provide better services to the public and one of the new enhancements offered by the screening services is to increase the screening age limit from 64 to 70 [26]. This effectively increases the number of screening episodes and results in augmenting the need for effective use of the already stretched NHS resources. All the aforementioned factors underline the need to increase the breast screening attendance.

III. Solution Proposed

To address these challenges, a set of protocols were developed as part of the ongoing research. The protocols are based on two components: 1) machine learning algorithms for knowledge creation; and 2) health informatics for knowledge sharing. This paper elaborates on how the prediction-based knowledge was created through a machine learning algorithm. Machine learning [Artificial Intelligence (AI)-based algorithm] was implemented through the creation of a prototype software based on open source technologies. The prototype software was automated to produce the preprocessed data and eventually normalize the data for neural network (AI) assimilation. These activities were performed sequentially without human involvement for repeatability, reliability, and accuracy.

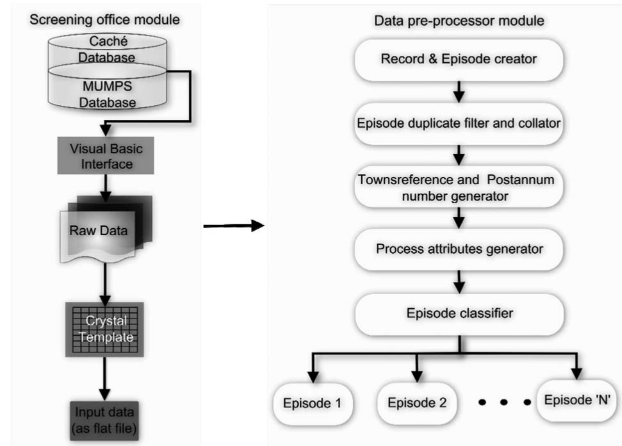

Fig. 1. Data filtering, preparation, and preprocessing.

The AI-based neural network incorporates all additional transformations that occurred within the screening process (including the change in the screening upper age limit). The prototype framework was called JAABS—Java-based attendance prediction by AI for breast screening. The prototype combines the demographic data pertaining to the nonattending women and information related to their family physician as a package. This package then triggers the generation of an electronic message based on the Health Level 7 (HL7) standards and utilizes web services as the message delivering technology. This paper focuses on the machine learning techniques used within the prototype and subsequent testing of the algorithm for its prediction accuracy.

A. Data Preprocessing Module

The prototype was constructed using two main modules: 1) data preprocessing module; and 2) AI module. The data preprocessing module (see Fig. 1) consists of "Screening office module" that accomplishes data extraction from the screening unit's database. The demography details for the three-year call/recall were downloaded (extraction date–Jan 2008) from the local health care authority's database. The downloading is affected via the health link network onto a standalone system within the breast screening unit. The historical data related to screening, appointments, and results pertaining to screening women are retained within the screening unit's "Massachusetts General Hospital Utility Multi-Programming System" (MUMPS) database. MUMPS, also known as the Oxford system, is one of the earliest programming languages used since the 1960s [27]. This language was extensively employed to write database applications explicitly for the healthcare domain.

```
Generate input data as flat file from "Crystal Report"
template
For every record
  Separate records for each woman
    Remove duplicate episodes
    Collate episodes into one record
  Generate townsend reference and post annum numbers
  Generate attributes
Classify and save record into their respective episode
groups
End
```

Pseudo-code 1. Pseudo-code for filtering raw data and preprocessing it to generate predictor attributes and classify them based on their episode details.

TABLE I
THIRTEEN-YEAR DATASET DETAILS

Description	Number of records
Total valid women's record	159,412
Number of records deleted due to multiple entries	15,778
Records with missing values	9,799
CR template output records	540,539

The MUMPS database is based on the disk operating system (DOS) and employs character-based user interface for database interrogation [27]. The cumbersome DOS-based system is prone to erroneous data entry and hence warranted a change in the system. A new software package, the National Breast Screening Computer System (NBSS), was developed in 2002–2003 to address these issues [28]. This NBSS consists of a Visual Basic (VB) front end connected to a "Caché" database which is seamlessly integrated with the MUMPS database [29]. Due to the aforementioned factors, an unstable environment, thus, resulted in considerable complexities during data extraction for the current research. The screening office module (see Fig. 1) is executed with the existing software programs available in the breast screening office.

The VB front end made data extraction straightforward from the MUMPS database through Structured Query Language (SQL) queries directed at the Caché database. Currently, the breast screening office is employing "Crystal Report" (CR) as part of the NBSS to generate reports for all the screening activities, including screening, administration, invitation, etc. Part of the data preprocessing was implemented through the CR software. The screening unit had earlier indicated that the routine functioning of the screening office should not be affected during the data extraction process.

Hence, prior to data extraction, a CR template was created to reflect the format of the data to be exported (see pseudo-code 1). This template was used to export the data as a flat file to negate any system instability. All the screening units around the country were expected to have some form of minimum facility for creating datasets in a flat file format. Coupled with this, a need for a low overhead on the existing IT system and minimum additional complexities was considered as fundamental for the prototype. All the aforementioned rationale strengthened the need for adopting a compromised strategy that exports data as a flat file, so that the mode of data transfer can be standardized across the country with minimum or no interrogation with the screening database.

The SQL query generated details for all the women in as many records, pertaining to the demography and episodes. The demographic data were incomplete and only the first record of a particular woman had the complete dataset and the remaining records of the women corresponded to the historical episode details (see Table I). The women's address and name were excluded from the study to address data protection and maintain anonymity. In spite of its necessity for the messaging module, the complete dataset was generated without the personal information of the screening women. The post code of the women is indispensable for the current study, as it generates the important predictor variable in the form of Townsend's reference (Townsend deprivation score denotes the socioeconomic status of a given postcode) and post annum number.

To address this without compromising the research work, variables related to postcode, such as the Townsend score, post annum (post annum is an arbitrary number associated with the women's postcode) and screening distance, were all processed to generate categorical variables within the screening unit and then the data were ported to the AI module. The individual women were identified by their SX number (pseudo-anonymised unique identifier). The AI module generated the attendance prediction, which formed the core of the knowledge transfer. The recipient of the knowledge transfer is the woman's family physician; hence, family physician information in the form of surname, surgery address, and postcode was later collated for sending the HL7-based message.

```
For each episode group
  Normalize data for AI module
  Generate networks (BPNN and RBFN) and train
  For each network
              Validate data
    Test data
    Generate screening attendance prediction
  Collate the best and save output with women's detail
End
```

Pseudo-code 2. Pseudo-code for the AI module and results collation for the final output

One "Record" object was associated with one or more "Episode" objects (see Fig. 2). The gaps in the demographic record have to be filled and the episode details were associated with the women's demographic data. Exhaustive analyses of the data indicated that the CR report had duplicate episode details and are to be removed before further processing can be implemented (see Table I). Each record read from the CR report has to be first partitioned into episode details and stored as "Episode" objects. They are finally collated and associated with the women's demographic details (represented as "Record" object). In addition to this, all the records have to be automatically validated. The earlier work by Arochena had identified all the contributing predictor attributes through comprehensive

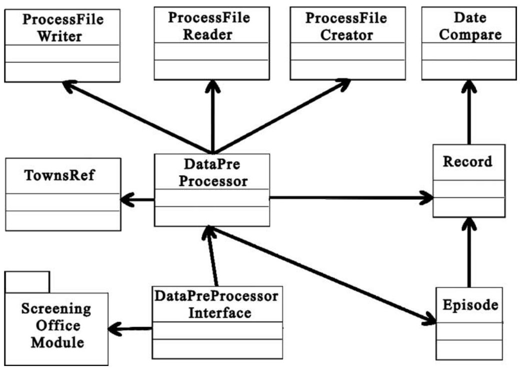

Fig. 2. UML class diagram for data preprocessing module (with I/O processing submodule).

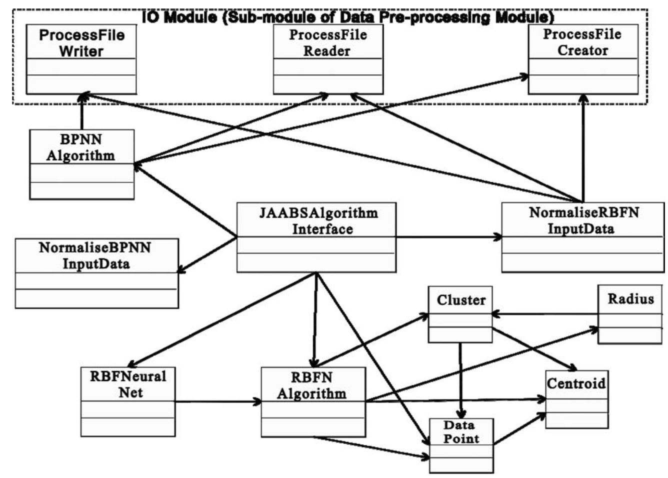

Fig. 3. UML class diagram of JAABS algorithm showing back propagation-based neural network and radial-basis function-based neural.

TABLE II
DATASET SPREAD ACROSS THE EPISODES AND ITS TRI-FURCATED DATA

Episode number	Total records	Train set	Valid set	Test set
Episode 1	23,277	4653	4708	13916
Episode 2	33,765	6838	6734	20193
Episode 3	29497	5868	5891	17738
Episode 4	43584	8792	8839	25953
Episode 5	26669	5340	5338	15991
Episode 6	2366	473	485	1408
Episode 7	238	36	39	163
Episode 8	16	3	3	10

statistical analyses [30]. After generating the required attributes, the preprocessor module classifies the "Record" objects based on the number of "Episode" objects it contains (see Fig. 2). This dataset was then written as an in-process flat file for reference. All errors generated during the execution of the preprocessing module are written in a log (error) and is also saved as a flat file for future reference.

The data preprocessing module identified episodes with missing data and removed them from the study. In total 2% (9 799) were removed as records with missing data (see Table I). It further deleted almost 3% (15 778) of the total records due to duplicate entries. The valid records constituted 86% (159 412) of the extracted dataset; on an average, each record had 3.2 episodes. Table II depicts the spread of data for each episode. The highest number of records was reached for the fourth episode. The first to fifth episodes had an average of 31 000 records. For the remaining episodes (sixth, seventh, and eighth) the average is only 800 records. This might have a significant impact on the actual prediction capacity of the JAABS algorithm for these episodes.

B. AI Module

JAABS is the new algorithm designed and developed in a JAVA environment. As the design process was based on more of an evolutionary type, a modular design strategy was selected. This assists in parallel development of the implementation and also enables testing as modules rather than as one single monolithic program. The modular design also ensured that any additions or changes happening within the screening unit's business logic can be implemented without affecting the other modules (see pseudo-code 2.). The "AI Module" encompasses the data normalizer; the neural networks; and the results collator (see Fig. 3). The Java-based algorithm implements two different neural networks: feed-forward back-propagation neural network (BPNN) and radial basis function neural network (RBFN).

The neural network algorithm requires the input data vector classified as binary values; hence, the input data are normalized. The input data in the RBFN are first passed through a radial basis function algorithm, to identify the clusters and assign a radius for cluster classification. These cluster centers are calculated and the real-time data are checked against these established cluster centers. Once the distance is calculated, the input dataset is then associated with its nearest cluster. These data then trigger a neural network for performing the prediction on attendance. Each episode has a different set of predictor attributes; hence, each episode is fed through separate neural networks that were trained with their respective training dataset.

The results module collects the collated prediction for each episode and submits it to a "Pooler" based classifier (see Fig. 4). The "Pooler" finds the best prediction for the given episode and generates the final prediction output based on the confidence value of the prediction. This is fed into the prediction result collator for all the input (women) based on each episode. The consolidated result is used to generate the nonattendance list and written as a flat file for processing by the "messaging module" for message generation. The final output is associated with the women's SX number so that general physician details can be added for knowledge sharing and to initiate physician intervention.

IV. ANALYSES

The predictor attributes (PA: post annum is an arbitrary number associated with the women's postcode, TS: townsend deprivation score denotes the socioeconomic status of a given postcode, AttBin: previous episode's attendance, NumTest: number of tests in the previous episodes, Cancer: denotes if cancer was diagnosed in previous episodes, FP: false positive in previous

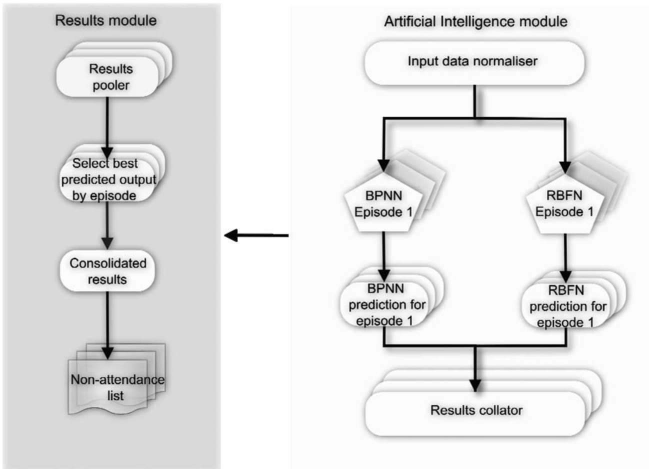

Fig. 4. Machine learning algorithm containing artificial intelligence and results module.

TABLE III
PREDICTOR ATTRIBUTES AND THEIR ASSOCIATION TO THE SCREENING ATTENDANCE EPISODE WISE

Independent variables	Epi1	Epi2	Epi3	Epi4	Epi5	Epi6	Epi7	Epi8
PA	❖	❖	❖	❖	❖	✓	✓	✓
TS	❖	❖	❖	❖	❖	❖	❖	✓
AttBin		✓	✓	✓	✓	✓	✓	✓
NumTest		✓	✓	✓	✓	✓	✓	✓
Cancer		○	○	❖	○	○	○	
FP		○	○	○	○	○	○	
HFP			○	○	○	○	○	✓
HC			○	○	○	○		
AttTypeBin	✓	✓	✓	✓	✓	✓	✓	✓
AgeBand	❖	❖	○	○	○	○		✓
Slip		✓	✓	✓	✓	✓	✓	✓
ScrDist	○	○	○	○	○	❖	✓	

✓ Association more than 0.2
❖ Association more than 0.1 and less than 0.2
○ Association more than 0.0 and less than 0.1
No association is left blank

TABLE IV
ROC FOR ALL EPISODES—AIATT AND JAABS (JAVA AND CLEMENTINE)

AI-ATT- Clementine (version 5)					
AIATT	ACC	NPV	PPV	SPC	SEN
Episode 1	67.01	20.45	87.48	41.81	71.43
Episode 2	87.76	56.1	92.85	58.91	93.14
Episode 3	86.49	50.54	92.91	55.99	91.32
Episode 4	81.65	41.26	92.51	64.59	85.42
Avg. for 4 Episodes	80.73	42.09	91.44	55.33	85.33
JAABS- Java					
JAABS	ACC	NPV	PPV	SPC	SEN
Episode 1	67.29	42.07	76.71	40.22	78.05
Episode 2	69.38	47.65	77.87	45.66	79.22
Episode 3	69.95	39.45	76.46	26.29	85.59
Episode 4	79.17	39.25	87.06	37.37	87.93
Episode 5	76.23	51.61	83.84	49.64	84.89
Episode 6	57.79	46.51	64.77	44.92	66.21
Episode 7	51.39	30.02	76.53	60.05	48.18
Avg. for 4 Episodes	71.45	42.11	79.53	37.39	82.7
Average	67.31	42.37	77.61	43.45	75.72
JAABS-Clementine (version 12)					
JAABS	ACC	NPV	PPV	SPC	SEN
Episode 1	68.16	52.58	69.35	11.57	95.04
Episode 2	79.61	74.59	81.33	57.93	90.28
Episode 3	81.24	72.56	83.86	57.63	90.99
Episode 4	85.73	74.91	88.45	62	93.34
Episode 5	80.81	74.43	82.56	53.88	92.18
Episode 6	67.88	63.8	70.36	56.7	76.16
Episode 7	78.99	86.49	77.61	41.56	96.89
Avg. for 4 Episodes	78.68	68.66	80.75	47.28	92.41
Average	77.49	71.34	79.08	48.75	90.7

episodes, HFP: history of false positive, HC: history of cancer, AttTypeBin: type of attendance like first or later episodes, AgeBand: age categories, Slip: difference in days between screening appointment and actual screening date, ScrDist: distance traveled by the women for getting a mammogram) were initially verified for their association with the screening attendance (see Table III). The variables, being categorical, were analyzed through parameters such as Lambda, Uncertainty, Phi (), Crammer's V, and Contingency (confidence level at 95%).

These tests for association were conducted for establishing some kind of linear relationship between the dependent and independent variables. Even though an association was not strong, it was used only to establish some form of relationship between the variables. This was used as an indication and as a first step for resolving the real problem space which is multispatial. This strategy assisted in filtering out the nonparticipating attributes and to reduce the introduction of background noise.

Episode 1 lacked the historical variables and had to rely only on demographic details. The rest of the episodes have both the demographic and historical attributes as predictors; especially the new attribute in the form of screening distance was found to increase the prediction efficiency for all the episodes. The JAABS algorithm and its predictor attributes were compared with its predecessor [AI-based attendance prediction algorithm(AI-ATT)] for validation [30]. The AI-ATT algorithm was developed in a visual modeling environment—Clementine [30]. This off-the-shelf software assisted in designing and implementing the algorithm rapidly, but created new functional challenges such as the need for licensing the software for all the screening units, specialist requirement for running the algorithm, as it was not automated, and is based on outdated data and semantics (1989–2001) to name just a few.

AI-ATT provided a base line for comparison and a reference for validating the JAABS algorithm. To make the validation more up-to-date, the same dataset that was applied to the JAABS algorithm was also tested on Clementine (version 12.0). The dataset was trifurcated into training, validating, and test sets (see Table II). The training set contained equal numbers of women categorized as attendees and nonattendees. The validating set contained data that were never exposed during the training and contained an equal number of attendees and nonattendees. The test set contained skewed data, where nonattendees were only a small proportion. This ensures that the test set reflects the realtime dataset that would also be skewed (less nonattendees). The JAABS algorithm was tested with the complete set of episodes after appropriate training and validation.

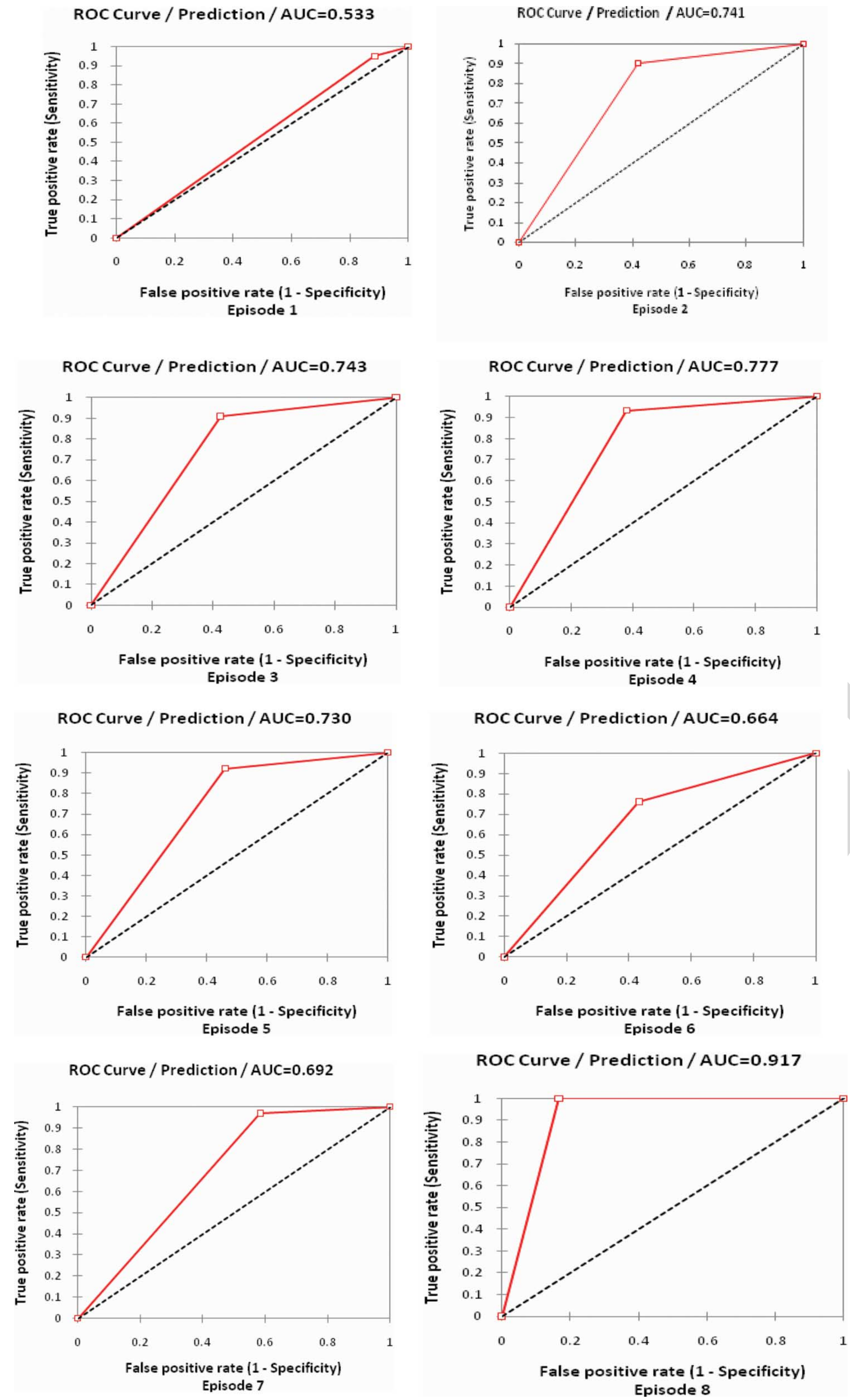

Fig. 5. ROC curve for Episodes one to eight for the machine learning algorithm.

The receiver operator characteristics (ROC) are summarized in Table IV (ACC: accuracy, NPV: negative predictive value, PPV: positive predictive value, SPC: specificity, SEN: sensitivity). The algorithm's final prediction of the screening attendance was based on a polling strategy that relies on the prediction confidence. The accuracy of the algorithm was around 68% for the first three episodes. Episode 4 had the maximum accuracy at 79%, closely followed by the fifth episode. The accuracies of the sixth and seventh episodes were lowest (57% and 51%, respectively). The NPV was the maximum at 51% for the fifth episode. The rest of the episodes had NPV values between 41% and 47%.

Episode 7 had the lowest NPV (30%). These lower NPVs were expected as the proportion of nonattendees was lesser in the test set (unbalanced). The PPVs for the fourth and fifth episodes were higher between 83% and 87%. The remaining episodes had values in the seventies range, except for the sixth episode where it was 64%. Specificity was highest for the seventh episode at 60%, but this may not be a true indicator as this episode had only 238 records in total. The next highest value was in the fifth episode at 49%. Episodes 1, 2, and 6 had values between 40% and 45%. Episodes 3 and 4 had lower values at 26% and 37%, respectively. The sensitivity was around 80% for the first four episodes, peaking at 85% for Episode 3. The higher the training set of records, the higher the sensitivity values. Since the previous algorithm (AI-ATT) had only four episodes, the averages for the first four episodes were used for comparing the JAABS and AI-ATT algorithms. The same set of attributes, when presented to commercial software (Clementine), generated improved results (see Table IV).

The first three episodes show an almost 10% increase in accuracy. Similarly, the later episodes (Episodes 4 and 5) when predicted by the JAABS–Clementine model, on average,do 6% better than the JAABS–Java algorithm, whereas Episodes 6 and 7 illustrated the maximum difference in accuracy (10–27%); this shows that the commercial software performed better even with a reduced training dataset. The NPV was lowest for the first episode, but was double when compared to AI-ATT and nearly 10% more than JAABS (Java). The NPV for the rest of the episodes (second to fifth) was around 73%. The remainder (sixth and seventh) were at 63% and 86%, respectively. The NPV is the metric that corresponds to the prediction of nonattendance and this was much better than that was achieved by the AI-ATT. Specificity is the next important measure and tests on Clementine showed promising results for all the episodes except for the first one.

The ROC curves for JAABS (Clementine) showed good prediction characteristics for all episodes except for Episode 1 (see Fig. 5). From the model's performance perspective, all these prediction characteristics were positive. The AI model proposed (JAABS—implemented in both Java and Clementine) was consistent and even outperformed the earlier model (AI-ATT) in many aspects. This could be attributed to the larger database and more complete attribute set and even the new predictor variable (screening distance) assisting in improving the algorithm's efficiency. The knowledge creation by applying AI (JAABS) is not only consistent, repeatable, and economical, but also ensures minimal human intervention. This is ideal for automating the whole process.

The proposed AI network (JAABS) for predicting screening nonattendance would be incorporated in a new breast screening software model that connects to the screening database to generate the screening batch. Based on the prediction, an automated message would be sent to the women's healthcare stakeholders (GPs, nurses, and other clinical specialists). These messages would be assimilated by the clinical system used by the stakeholders and would eventually flag the women as a nonattendee. When a woman's clinical record is opened, a flag/pop-up window would trigger opportunistic interventions that are aimed at educating the woman. This knowledge transfer would empower the woman to make an informed decision toward screening. This multistakeholder-based opportunistic intervention strategy would increase the overall breast screening attendance.

V. Conclusion

This paper discussed the details of how a machine learning-based prediction tool can be effectively applied to increase the breast cancer screening attendance. The need for a high degree of automation was highlighted to simplify the algorithm's adoption; such automation would also reduce overheads and make integration as seamless as possible [31]. From the model's performance perspective, all the prediction characteristics were positive. The machine learning-based AI model (JAABS—implemented in both Java and Clementine) proposed was consistent and even outperformed the earlier model (AI-ATT) in many aspects. The performance improvement could be attributed to the larger database, more complete attribute set and even the new predictor variable (screening distance). The knowledge creation by applying AI (JAABS) is not only reliable, repeatable, and economical, but also ensures minimal human intervention. There is still scope for improving the prediction efficiency and this can be achieved through better predictor attributes and/or improved machine learning techniques. The former would be difficult to achieve as the data source itself may not be available but the latter would be possible as better AI models, such as support vector machines, fuzzy logic, and genetic algorithms or a combination of these, would enable further investigation for increasing the efficiency.

Acknowledgment

The authors would like to thank J. Patnick CBE, Director, NHS Cancer Screening Programs (U.K.), for funding this research, Dr. M. Wallis, Consultant Radiologist, Cambridge Breast Unit team, and Margot Wheaton, Program Manager for the Warwickshire, Solihull and Coventry Breast Screening Service at Coventry and Warwickshire Hospital, for their excellent support and guidance throughout this research.

References

[1] American Cancer Society. (2010, Feb. 10). *Breast Cancer Facts & Figures 2009–2010* [Online]. Available: http://www.acsevents.org/downloads/STT/F861009_final%209-08-09.pdf.

[2] Cancer Research U.K. (2010, Feb. 10). *Breast Cancer—U.K. Mortality Statistics*. [Online]. Available: http://info.cancerresearchuk.org/cancerstats/types/breast/mortality/index.htm.

[3] NHS Breast Screening Programme—Cancer Screening Programmes Annual Review 2009. (2010, Feb. 10). [Online]. Available: http://www.cancerscreening.nhs.uk/breastscreen/publications/nhsbsp-annualreview2009.pdf.

[4] K. Turner, J. Wilson, and J. Gilbert, "Improving breast screening uptake: Persuading initial non-attenders to attend," *J. Med. Screening*, vol. 1, Q2 pp. 199–202, 1994.

[5] A. Majeed, R. Given-Wilson, and E. Smith, "Impact of follow up letters on non-attenders for breast screening: A general practice based study," *J. Med. Screening*, vol. 4, pp. 19–20, 1997.

[6] J. P. Sin and A. S. Leger, "Interventions to increase breast screening uptake: Do they make any difference?," *J. Med. Screening*, vol. 6, no. 1, pp. 170–181, 1999.

[7] Canadian Cancer Society. (2006). *Canadian Researchers Find Common Breast Cancer Chemotherapy Regime Inferior at Preventing Disease Recurrence* [Online]. Available: http://www.cancer.ca/Canadawide/About%20us/Media%20centre/CW-Media%20releases/CW2006/Canadian%20Researchers%20Find%20Common%20Breast%20Cancer%20Chemotherapy%20Regime%20Inferior%20at%20Preventing%20Disease%20Recurrence.aspx?sc_lang=en.

[8] Canadian Cancer Society. (2008, Mar. 22). *Canadian Cancer Statistics 2008* [Online]. Available: http://www.cancer.ca/Canada-wide/About%20cancer/Cancer%20statistics/~/media/CCS/Canada%20wide/Files%20List/English%20files%20heading/pdf%20not%20in%20publications%20section/Canadian%20Cancer%20Society%20Statistics%20PDF%202008_614137951.ashx.

[9] A. Oikonomou, S. A. Amin, R. N. G. Naguib, A. Todman, and H. Al-Omishy, "Breast self examination training through the use of multimedia: A prototype multimedia application," *IEEE Eng. Med. Biol. Soc.*, vol. 2, no. 21, pp. 295–298, 2003.

[10] B. V. Marcela, "The system does work," *J. Am. College Radiol.*, vol. 1, no. 6, pp. 438–440, 2004.

[11] L. Wyld, "Mammographic Breast Screening in Elderly Women," in *Management of Breast Cancer in Older Women*, part 3, M. W. Reed and R. A. Audisio, Eds. London, U.K.: Springer, 2010, ch. 9, pp. 127–142.

[12] R. G. Blanks, S. M. Moss, C. E. McGahan, M. J. Quinn, and P. J. Babb, "Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990–1998: Comparison of observed with predicted mortality," *BMJ*, vol. 321, no. 7262, pp. 665–669, 2000.

[13] S. S. Epstein, *The Politics of Cancer*. New York: Doubleday, 1979, pp. 537.

[14] G. Burton, *Alternative Medicine*. Washington, DC: Future Medicine Publishing, 1997.

[15] Cancer Research U.K. (2007, Jul. 14). *Cancer Incidence—U.K. Statistics* [Online]. Available: http://info.cancerresearchuk.org/cancerstats/incidence/index.htm

[16] P. Forest, *Breast Cancer Screening—A Report to the Health Ministers of England, Scotland, Wales and Northern Ireland*. London, U.K.: HMSO, 1986.

[17] Medicine net (2010 Feb. 18). *Breast Cancer* [Online]. Available: http://www.medicinenet.com/breast_cancer/page3.htm

[18] I. Pirjo, L. Kauhava, I. Parvinen, H. Helenius, and P. Klemi, "Customer fee and participation in breast cancer screening," *The Lancet*, vol. 358, p. 1425, 2001.

[19] S. H. Woolf, "The 2009 Breast Cancer Screening Recommendations of the US Preventive Services Task Force," *JAMA*, vol. 303, no. 2, pp. 162–163, 2010.

[20] American Cancer Society Inc., (2010, Feb. 18) *Cancer Reference Information* [Online]. Available: http://www.cancer.org/docroot/CRI/CRI_2_5x.asp?dt=5

[21] D. P. Weller and C. Campbell, "Uptake in cancer screening programmes: A priority in cancer control," *Brit. J. Cancer*, vol. 101, pp. 55–59, 2009.

[22] Y. Zheng, "Breast cancer detection with gabor features from digital mammograms," *Algorithms*, vol. 3, pp. 44–62, 2010.

[23] K. W. Eilbert, K. Carroll, J. Peach, S. Khatoon, I. Basnett, and N. McCulloch, "Approaches to improving breast screening uptake: Evidence and experience from Tower Hamlets," *Brit. J. Cancer*, vol. 101, no. 2, pp. 64–67, 2009.

[24] D. Schopper and C. de Wolf, "How effective are breast cancer screening programmes by mammography? Review of the current evidence," *Eur. J. Cancer.*, vol. 45, no. 11, pp. 1916–1923, Jul. 2009.

[25] E. S. Cassandra, "Breast cancer screening: Cultural beliefs and diverse populations," *Health Soc. work*, vol. 31, no. 1, pp. 36–43, 2006.

[26] NHS Cancer Screening Programmes. (2007, Apr.) *Disclosure of Audit Results in Cancer Screening Advice on Best Practice* (Cancer Screening Series 3), J. Patnick, Ed. [Online]. Available: http://www.cancerscreening.nhs.uk/publications/cs3.pdf

[27] K. Okane. (2005, Apr. 20). *Mumps Language Bioinformatic Database Resources* [Online]. Available: http://bioinformatics.org/forums/forum.php?forum_id=1035

[28] V. Baskaran, R. K. Bali, R. N. G. Naguib, and H. Arochena, "A Knowledge Management approach to increase uptake in a breast screening programme," presented at the IEEE 2nd Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) Int. Conf., Philippines, Mar. 2005.

[29] S. Tarver, K. Cronin-Cowan, and M. E. Wheaton, "A pilot's life for us," *Breast Cancer Res.*, vol. 6, suppl. 1, p. 52, 2004.

[30] H. E. Arochena, "Modelling and prediction of parameters affecting attendance to the NHS breast cancer screening programme," Ph.D. dissertation, Dept. Comp. Sci., Coventry Univ., Coventry, U.K., 2003.

[31] C. Bankhead, S. H. Richards, T. Peters, D. Sharp, R. Hobbs, J. Brown, L. Roberts, C. Tydeman, V. Redman, J. Formby, S. Wilson, and J. Austoker, "Improving attendance for breast screening among recent non-attenders: A randomised controlled trial of two interventions in primary care," *J. Med. Screening*, vol. 8, no. 2, pp. 99–105, 2001.

Vikraman Baskaran is currently an Assistant Professor at the School of Information Technology Management of Ryerson University, Toronto, ON, Canada. His research interests include finding a viable application of the KM paradigm in healthcare application. His special interest in developing HL7 messaging and health informatics has provided opportunities in excelling in these fields. His current activities include KM, e-health, artificial intelligence, and healthcare informatics. Q3

He is a member of the HL7 U.K. and Canada. Q4

Aziz Guergachi (M'xx) is currently an Associate Professor at the Ted Rogers School of Information Technology Management of Ryerson University, Toronto, ON, Canada. Prior to becoming part of the Ryerson community, he was involved in the development of a large software system for trade promotion management and collaborative sales forecasting. His current research interests include advanced system modeling and machine learning with applications to business management and engineering systems.

He is the recipient of the New Opportunities Award of the Canada Foundation for Innovation and currently runs a research laboratory for advanced systems modeling. Q5

Rajeev K. Bali (SM'xx) is currently a Reader in Healthcare Knowledge Management at Coventry University, U.K. His main research interests include clinical and healthcare knowledge management (from both technical and organisational perspectives). He has published peer-reviewed journals and is the an author/editor of several textbooks on healthcare knowledge management.

He serves on various editorial boards and conference committees and is regularly invited to deliver presentations and speeches internationally. Q6

Raouf N. G. Naguib (SM'xx) is currently a Professor of Biomedical Computing and Head of BIOCORE, Coventry, U.K. Prior to this appointment, he was a Lecturer at Newcastle University, Newcastle Upon Tyne, U.K. He has published more than 240 journals and conference papers and reports in many aspects of biomedical and digital signal processing, image processing, artificial intelligence, and evolutionary computation in cancer research. Q7

He was awarded the Fulbright Cancer Fellowship in 1995–1996 when he carried out research at the University of Hawaii, Mãnoa, on the applications of artificial neural networks in breast cancer diagnosis and prognosis. He is a member of several national and international research committees and boards.

QUERIES

Q1. Author: Please check whether the edits made in the sentence "This large percentage of nonattendance not only..." retain your intended sense.
Q2. Author: Refs. [5], [6], [8], [12], [14], [15], [16], and [23] are not cited in the text. Please check and provide citations.
Q3. Author: Please provide the expansion of KM.
Q4. Author: Please provide the educational details of all the authors.
Q5. Author: Please provide the year in which Aziz Guergachi became "Member" of the IEEE.
Q6. Author: Please provide the year in which Rajeev K Bali became "Senior Member" of the IEEE.
Q7. Author: Please provide the year in which Raouf N. G. Naguib became "Senior Member" of the IEEE.

Predicting Breast Screening Attendance Using Machine Learning Techniques

Vikraman Baskaran, Aziz Guergachi, *Member, IEEE*, Rajeev K. Bali, *Senior Member, IEEE*, and Raouf N. G. Naguib, *Senior Member, IEEE*

***Abstract*—Machine learning-based prediction has been effectively applied for many healthcare applications. Predicting breast screening attendance using machine learning (prior to the actual mammogram) is a new field. This paper presents new predictor attributes for such an algorithm. It describes a new hybrid algorithm that relies on back-propagation and radial basis function-based neural networks for prediction. The algorithm has been developed in an open source-based environment. The algorithm was tested on a 13-year dataset (1995–2008). This paper compares the algorithm and validates its accuracy and efficiency with different platforms. Nearly 80% accuracy and 88% positive predictive value and sensitivity were recorded for the algorithm. The results were encouraging; 40–50% of negative predictive value and specificity warrant further work. Preliminary results were promising and provided ample amount of reasons for testing the algorithm on a larger scale.**

***Index Terms*—Breast screening, cancer, machine learning, neural networks, prediction, screening attendance.**

I. Introduction

BREAST cancer is the most common cancer for women in North America [1]. In the U.K., over 40 000 women are being diagnosed with breast cancer each year [2], [3]. Mortality due to breast cancer is also one of the highest in the world [1], [4], and is the second highest of all cancers in the Canada [7]. Breast cancer should ideally be diagnosed at the earlier stages of its development to considerably reduce mortality. Possible treatments include removing or destroying the cancer cells to avoid the spread of the affected cells. Breast self-examination is an effective and noninvasive type of checking for any lumps in the breast tissue. Unfortunately, this greatly depends on the size of the lump, technique, and experience in carrying out a self-examination procedure by a woman [9]. An ultrasound test, examining breast tissue using sound waves, can be utilized to detect lumps but this is usually suited for women aged below 35 owing to the higher density of breast tissue [1]. Having a tissue biopsy via a fine needle aspiration or an excision is often used to examine the cells histopathologically and to diagnose if the growth, lump, is benign or cancerous. These investigations are mostly employed in treatments or post-treatment examination and as second rung diagnostic confirmation methods [10]. Performing a computed tomography or an MRI scan would result in a thorough examination of the breast tissue but these techniques are not favored due to reasons which include cost, needs preparation, noise, time, and images that may not be clear [10].

Mammography is a technique for detecting breast tissue lumps using a low dosage of X-ray. This technique can even detect a 3-mm-sized lump. The X-ray image of the breast tissue is captured and the image is thoroughly read by experienced radiologists and specialist mammogram readers [10]. Preliminary research suggests that women aged 50 and above are more susceptible to breast cancer; mammography is more suited to women in this age range due to the lower density of breast tissue [11]. Even though mammography has its critics—mainly due to its high rate of false positives and false negatives [13]—it has become the standard procedure for screening women by the NHS National Breast Screening Program in the U.K. [3], [15]. Mammography is the best and most viable tool for mass screening to detect cancer in the breast at an early stage [17]; however, the effectiveness of diagnosis through screening is directly dependent on the percentage of women attending the screening program [18]–[20]. The NHS Breast Screening Program, catering to the entire eligible women population, is funded by the Department of Health in the U.K. It covers 2.5 million women every year and detected nearly 16 500 cancers in the screened population for the year 2007–2008 [3]. Currently, the screening program routinely screens women between the ages 50 and 70.

Early breast cancer detection through screening is fundamental for increasing the efficacy of cancer treatment [11], [21]. Mammography has been accepted as the best and most economically viable tool for population screening [22]. Maximizing coverage for the target population is crucial for the success of such screening programs [11]. Currently, the breast cancer screening attendance rates are below expectations in many countries that have publicly funded healthcare programs [24]. This paper proposes a set of protocols to increase breast screening attendance for the U.K.'s NHS breast screening program. Based on this protocol, a new software prototype was created and tested. The prototype tests the prediction algorithm and shares the prediction results with multiple healthcare stakeholders for initiating opportunistic interventions on nonattendees. This prototype is a radical new idea that uses machine learning techniques for

Manuscript received April 11, 2010; revised October 12, 2010; accepted December 23, 2010. Date of publication; date of current version. This work was supported in part by the NHS Cancer Screening Programs, U.K.

V. Baskaran and A. Guergachi are with Ryerson University, TRSM, Toronto, ON M5B 2K3, Canada (e-mail: vikraman@ryerson.ca; a2guerga@ryerson.ca).

R. K. Bali is with the KARMAH Group, Health Design and Technology Institute (HDTI), Coventry University, Coventry University Technology Park, Coventry CV1 2TT, U.K. (e-mail: r.bali@coventry.ac.uk).

R. N. G. Naguib is with the Biomedical Computing and BIOCORE, HDTI, Coventry University, Coventry University Technology Park, Coventry CV1 2TT, U.K. (e-mail: r.naguib@coventry.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2010.2103954

predicting screening attendance and shares this knowledge by adopting the health informatics initiative of the NHS.

II. Challenge

The NHS Breast Screening Program Annual Review (2008) states that, out of invited women, only 74% attend the screening program [3]. This sizeable nonattendance could result in missed cancer detection for nearly 4 000 women (based on the cancer detection rate within screened women) [3]. This large percentage of nonattendance not only result in loss of life due to breast cancer but also result in loss of screening resources through costly imaging equipment laying idle, underutilization of specialist-imaging expertise, wasted screening slots, and so forth. Screening units are unable to arrange buffered attendees for the idle slots since the units do not know *a priori* which women will attend and which will not. In addition, there is a sizeable cost factor involved in sending repeat screening appointments letters to nonattending women.

Reasons for nonattendance may be largely attributed to disinterest in attending a mammography session, prior or current medical problems, and fear of X-rays [11], [24]. These reasons can be negated by proper education provided to women. Education has to be directed at explaining the advantages and importance of screening and assist in removing the sociocultural and personal barriers [25]. Other possible options include convenience in terms of time, place, and dates provided to women for encouraging their attendance.

In spite of the expedient measures provided to the women, nonattendance has been a grave concern for the NHS—National Screening Program. This scenario can be properly addressed if those women who may probably not attend a screening appointment can be identified in advance so that additional resources can be directed at interventions that can increase screening attendance.

A proposal enumerating the complete software solution is summarized at the end of Section IV. The National Screening Program has been constantly striving to provide better services to the public and one of the new enhancements offered by the screening services is to increase the screening age limit from 64 to 70 [26]. This effectively increases the number of screening episodes and results in augmenting the need for effective use of the already stretched NHS resources. All the aforementioned factors underline the need to increase the breast screening attendance.

III. Solution Proposed

To address these challenges, a set of protocols were developed as part of the ongoing research. The protocols are based on two components: 1) machine learning algorithms for knowledge creation; and 2) health informatics for knowledge sharing. This paper elaborates on how the prediction-based knowledge was created through a machine learning algorithm. Machine learning [Artificial Intelligence (AI)-based algorithm] was implemented through the creation of a prototype software based on open source technologies. The prototype software was automated to produce the preprocessed data and eventually normalize the data for neural network (AI) assimilation. These activities were performed sequentially without human involvement for repeatability, reliability, and accuracy.

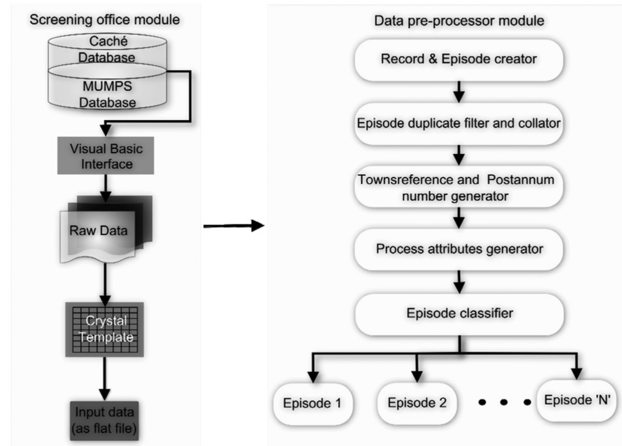

Fig. 1. Data filtering, preparation, and preprocessing.

The AI-based neural network incorporates all additional transformations that occurred within the screening process (including the change in the screening upper age limit). The prototype framework was called JAABS—Java-based attendance prediction by AI for breast screening. The prototype combines the demographic data pertaining to the nonattending women and information related to their family physician as a package. This package then triggers the generation of an electronic message based on the Health Level 7 (HL7) standards and utilizes web services as the message delivering technology. This paper focuses on the machine learning techniques used within the prototype and subsequent testing of the algorithm for its prediction accuracy.

A. Data Preprocessing Module

The prototype was constructed using two main modules: 1) data preprocessing module; and 2) AI module. The data preprocessing module (see Fig. 1) consists of "Screening office module" that accomplishes data extraction from the screening unit's database. The demography details for the three-year call/recall were downloaded (extraction date–Jan 2008) from the local health care authority's database. The downloading is affected via the health link network onto a standalone system within the breast screening unit. The historical data related to screening, appointments, and results pertaining to screening women are retained within the screening unit's "Massachusetts General Hospital Utility Multi-Programming System" (MUMPS) database. MUMPS, also known as the Oxford system, is one of the earliest programming languages used since the 1960s [27]. This language was extensively employed to write database applications explicitly for the healthcare domain.

```
Generate input data as flat file from "Crystal Report"
template
For every record
   Separate records for each woman
      Remove duplicate episodes
      Collate episodes into one record
   Generate townsend reference and post annum numbers
   Generate attributes
Classify and save record into their respective episode
groups
End
```

Pseudo-code 1. Pseudo-code for filtering raw data and preprocessing it to generate predictor attributes and classify them based on their episode details.

TABLE I
THIRTEEN-YEAR DATASET DETAILS

Description	Number of records
Total valid women's record	159,412
Number of records deleted due to multiple entries	15,778
Records with missing values	9,799
CR template output records	540,539

The MUMPS database is based on the disk operating system (DOS) and employs character-based user interface for database interrogation [27]. The cumbersome DOS-based system is prone to erroneous data entry and hence warranted a change in the system. A new software package, the National Breast Screening Computer System (NBSS), was developed in 2002–2003 to address these issues [28]. This NBSS consists of a Visual Basic (VB) front end connected to a "Caché" database which is seamlessly integrated with the MUMPS database [29]. Due to the aforementioned factors, an unstable environment, thus, resulted in considerable complexities during data extraction for the current research. The screening office module (see Fig. 1) is executed with the existing software programs available in the breast screening office.

The VB front end made data extraction straightforward from the MUMPS database through Structured Query Language (SQL) queries directed at the Caché database. Currently, the breast screening office is employing "Crystal Report" (CR) as part of the NBSS to generate reports for all the screening activities, including screening, administration, invitation, etc. Part of the data preprocessing was implemented through the CR software. The screening unit had earlier indicated that the routine functioning of the screening office should not be affected during the data extraction process.

Hence, prior to data extraction, a CR template was created to reflect the format of the data to be exported (see pseudo-code 1). This template was used to export the data as a flat file to negate any system instability. All the screening units around the country were expected to have some form of minimum facility for creating datasets in a flat file format. Coupled with this, a need for a low overhead on the existing IT system and minimum additional complexities was considered as fundamental for the prototype. All the aforementioned rationale strengthened the need for adopting a compromised strategy that exports data as a flat file, so that the mode of data transfer can be standardized across the country with minimum or no interrogation with the screening database.

The SQL query generated details for all the women in as many records, pertaining to the demography and episodes. The demographic data were incomplete and only the first record of a particular woman had the complete dataset and the remaining records of the women corresponded to the historical episode details (see Table I). The women's address and name were excluded from the study to address data protection and maintain anonymity. In spite of its necessity for the messaging module, the complete dataset was generated without the personal information of the screening women. The post code of the women is indispensable for the current study, as it generates the important predictor variable in the form of Townsend's reference (Townsend deprivation score denotes the socioeconomic status of a given postcode) and post annum number.

To address this without compromising the research work, variables related to postcode, such as the Townsend score, post annum (post annum is an arbitrary number associated with the women's postcode) and screening distance, were all processed to generate categorical variables within the screening unit and then the data were ported to the AI module. The individual women were identified by their SX number (pseudo-anonymised unique identifier). The AI module generated the attendance prediction, which formed the core of the knowledge transfer. The recipient of the knowledge transfer is the woman's family physician; hence, family physician information in the form of surname, surgery address, and postcode was later collated for sending the HL7-based message.

```
For each episode group
   Normalize data for AI module
   Generate networks (BPNN and RBFN) and train
   For each network
                  Validate data
      Test data
      Generate screening attendance prediction
   Collate the best and save output with women's detail
End
```

Pseudo-code 2. Pseudo-code for the AI module and results collation for the final output

One "Record" object was associated with one or more "Episode" objects (see Fig. 2). The gaps in the demographic record have to be filled and the episode details were associated with the women's demographic data. Exhaustive analyses of the data indicated that the CR report had duplicate episode details and are to be removed before further processing can be implemented (see Table I). Each record read from the CR report has to be first partitioned into episode details and stored as "Episode" objects. They are finally collated and associated with the women's demographic details (represented as "Record" object). In addition to this, all the records have to be automatically validated. The earlier work by Arochena had identified all the contributing predictor attributes through comprehensive

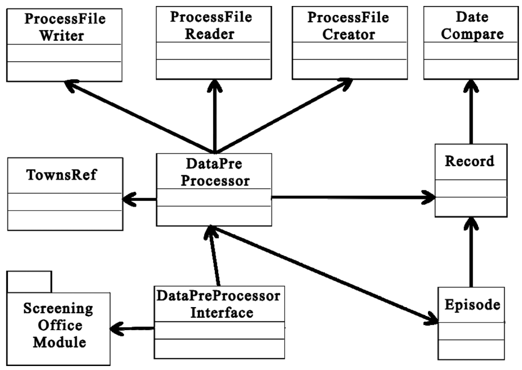

Fig. 2. UML class diagram for data preprocessing module (with I/O processing submodule).

TABLE II
DATASET SPREAD ACROSS THE EPISODES AND ITS TRI-FURCATED DATA

Episode number	Total records	Train set	Valid set	Test set
Episode 1	23,277	4653	4708	13916
Episode 2	33,765	6838	6734	20193
Episode 3	29497	5868	5891	17738
Episode 4	43584	8792	8839	25953
Episode 5	26669	5340	5338	15991
Episode 6	2366	473	485	1408
Episode 7	238	36	39	163
Episode 8	16	3	3	10

statistical analyses [30]. After generating the required attributes, the preprocessor module classifies the "Record" objects based on the number of "Episode" objects it contains (see Fig. 2). This dataset was then written as an in-process flat file for reference. All errors generated during the execution of the preprocessing module are written in a log (error) and is also saved as a flat file for future reference.

The data preprocessing module identified episodes with missing data and removed them from the study. In total 2% (9 799) were removed as records with missing data (see Table I). It further deleted almost 3% (15 778) of the total records due to duplicate entries. The valid records constituted 86% (159 412) of the extracted dataset; on an average, each record had 3.2 episodes. Table II depicts the spread of data for each episode. The highest number of records was reached for the fourth episode. The first to fifth episodes had an average of 31 000 records. For the remaining episodes (sixth, seventh, and eighth) the average is only 800 records. This might have a significant impact on the actual prediction capacity of the JAABS algorithm for these episodes.

B. AI Module

JAABS is the new algorithm designed and developed in a JAVA environment. As the design process was based on more of an evolutionary type, a modular design strategy was selected. This assists in parallel development of the implementation and also enables testing as modules rather than as one single monolithic program. The modular design also ensured that any additions or changes happening within the screening unit's business logic can be implemented without affecting the other modules (see pseudo-code 2.). The "AI Module" encompasses the data normalizer; the neural networks; and the results collator (see Fig. 3). The Java-based algorithm implements two different neural networks: feed-forward back-propagation neural network (BPNN) and radial basis function neural network (RBFN).

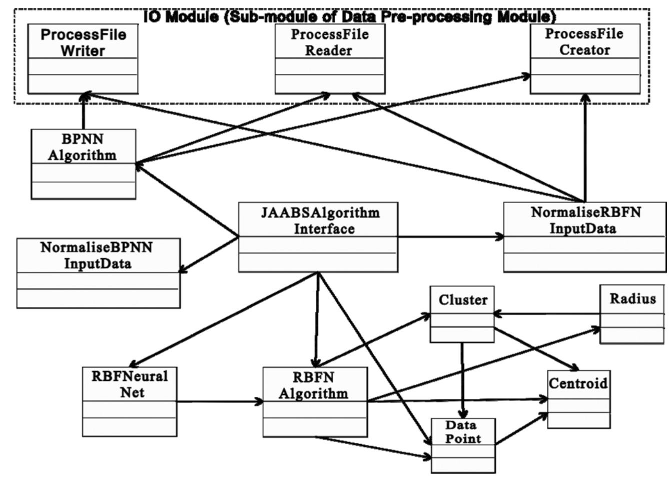

Fig. 3. UML class diagram of JAABS algorithm showing back propagation-based neural network and radial-basis function-based neural.

The neural network algorithm requires the input data vector classified as binary values; hence, the input data are normalized. The input data in the RBFN are first passed through a radial basis function algorithm, to identify the clusters and assign a radius for cluster classification. These cluster centers are calculated and the real-time data are checked against these established cluster centers. Once the distance is calculated, the input dataset is then associated with its nearest cluster. These data then trigger a neural network for performing the prediction on attendance. Each episode has a different set of predictor attributes; hence, each episode is fed through separate neural networks that were trained with their respective training dataset.

The results module collects the collated prediction for each episode and submits it to a "Pooler" based classifier (see Fig. 4). The "Pooler" finds the best prediction for the given episode and generates the final prediction output based on the confidence value of the prediction. This is fed into the prediction result collator for all the input (women) based on each episode. The consolidated result is used to generate the nonattendance list and written as a flat file for processing by the "messaging module" for message generation. The final output is associated with the women's SX number so that general physician details can be added for knowledge sharing and to initiate physician intervention.

IV. ANALYSES

The predictor attributes (PA: post annum is an arbitrary number associated with the women's postcode, TS: townsend deprivation score denotes the socioeconomic status of a given postcode, AttBin: previous episode's attendance, NumTest: number of tests in the previous episodes, Cancer: denotes if cancer was diagnosed in previous episodes, FP: false positive in previous

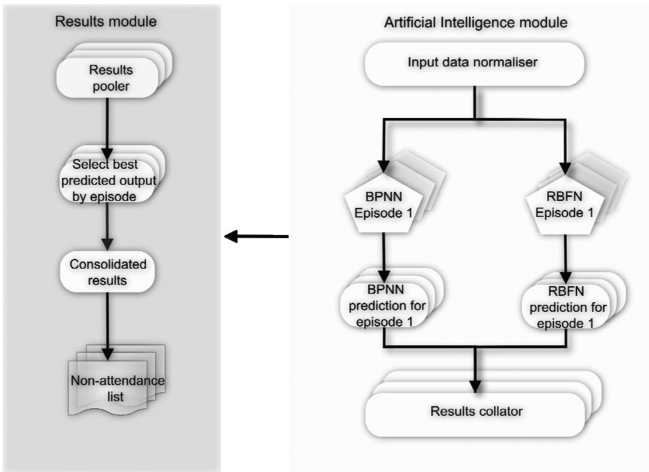

Fig. 4. Machine learning algorithm containing artificial intelligence and results module.

TABLE III
PREDICTOR ATTRIBUTES AND THEIR ASSOCIATION TO THE SCREENING ATTENDANCE EPISODE WISE

Independent variables	Epi1	Epi2	Epi3	Epi4	Epi5	Epi6	Epi7	Epi8
PA	❖	❖	❖	❖	❖	✓	✓	✓
TS	❖	❖	❖	❖	❖	❖	❖	✓
AttBin		✓	✓	✓	✓	✓	✓	✓
NumTest		✓	✓	✓	✓	✓	✓	✓
Cancer		○	○	❖	○	○	○	
FP		○	○	○	○	○	○	
HFP			○	○	○	○	○	✓
HC			○	○	○	○		
AttTypeBin	✓	✓	✓	✓	✓	✓	✓	✓
AgeBand	❖	❖	○	○	○	○		✓
Slip		✓	✓	✓	✓	✓	✓	✓
ScrDist	○	○	○	○	○	❖	✓	

✓ Association more than 0.2
❖ Association more than 0.1 and less than 0.2
○ Association more than 0.0 and less than 0.1
No association is left blank

episodes, HFP: history of false positive, HC: history of cancer, AttTypeBin: type of attendance like first or later episodes, AgeBand: age categories, Slip: difference in days between screening appointment and actual screening date, ScrDist: distance traveled by the women for getting a mammogram) were initially verified for their association with the screening attendance (see Table III). The variables, being categorical, were analyzed through parameters such as Lambda, Uncertainty, Phi (), Crammer's V, and Contingency (confidence level at 95%).

These tests for association were conducted for establishing some kind of linear relationship between the dependent and independent variables. Even though an association was not strong, it was used only to establish some form of relationship between the variables. This was used as an indication and as a first step for resolving the real problem space which is multispatial. This strategy assisted in filtering out the nonparticipating attributes and to reduce the introduction of background noise.

Episode 1 lacked the historical variables and had to rely only on demographic details. The rest of the episodes have both the demographic and historical attributes as predictors; especially the new attribute in the form of screening distance was found to increase the prediction efficiency for all the episodes. The JAABS algorithm and its predictor attributes were compared with its predecessor [AI-based attendance prediction algorithm(AI-ATT)] for validation [30]. The AI-ATT algorithm was developed in a visual modeling environment—Clementine [30]. This off-the-shelf software assisted in designing and implementing the algorithm rapidly, but created new functional challenges such as the need for licensing the software for all the screening units, specialist requirement for running the algorithm, as it was not automated, and is based on outdated data and semantics (1989–2001) to name just a few.

TABLE IV
ROC FOR ALL EPISODES—AIATT AND JAABS (JAVA AND CLEMENTINE)

AI-ATT- Clementine (version 5)					
AIATT	ACC	NPV	PPV	SPC	SEN
Episode 1	67.01	20.45	87.48	41.81	71.43
Episode 2	87.76	56.1	92.85	58.91	93.14
Episode 3	86.49	50.54	92.91	55.99	91.32
Episode 4	81.65	41.26	92.51	64.59	85.42
Avg. for 4 Episodes	80.73	42.09	91.44	55.33	85.33
JAABS- Java					
JAABS	ACC	NPV	PPV	SPC	SEN
Episode 1	67.29	42.07	76.71	40.22	78.05
Episode 2	69.38	47.65	77.87	45.66	79.22
Episode 3	69.95	39.45	76.46	26.29	85.59
Episode 4	79.17	39.25	87.06	37.37	87.93
Episode 5	76.23	51.61	83.84	49.64	84.89
Episode 6	57.79	46.51	64.77	44.92	66.21
Episode 7	51.39	30.02	76.53	60.05	48.18
Avg. for 4 Episodes	71.45	42.11	79.53	37.39	82.7
Average	67.31	42.37	77.61	43.45	75.72
JAABS-Clementine (version 12)					
JAABS	ACC	NPV	PPV	SPC	SEN
Episode 1	68.16	52.58	69.35	11.57	95.04
Episode 2	79.61	74.59	81.33	57.93	90.28
Episode 3	81.24	72.56	83.86	57.63	90.99
Episode 4	85.73	74.91	88.45	62	93.34
Episode 5	80.81	74.43	82.56	53.88	92.18
Episode 6	67.88	63.8	70.36	56.7	76.16
Episode 7	78.99	86.49	77.61	41.56	96.89
Avg. for 4 Episodes	78.68	68.66	80.75	47.28	92.41
Average	77.49	71.34	79.08	48.75	90.7

AI-ATT provided a base line for comparison and a reference for validating the JAABS algorithm. To make the validation more up-to-date, the same dataset that was applied to the JAABS algorithm was also tested on Clementine (version 12.0). The dataset was trifurcated into training, validating, and test sets (see Table II). The training set contained equal numbers of women categorized as attendees and nonattendees. The validating set contained data that were never exposed during the training and contained an equal number of attendees and nonattendees. The test set contained skewed data, where nonattendees were only a small proportion. This ensures that the test set reflects the real-time dataset that would also be skewed (less nonattendees). The JAABS algorithm was tested with the complete set of episodes after appropriate training and validation.

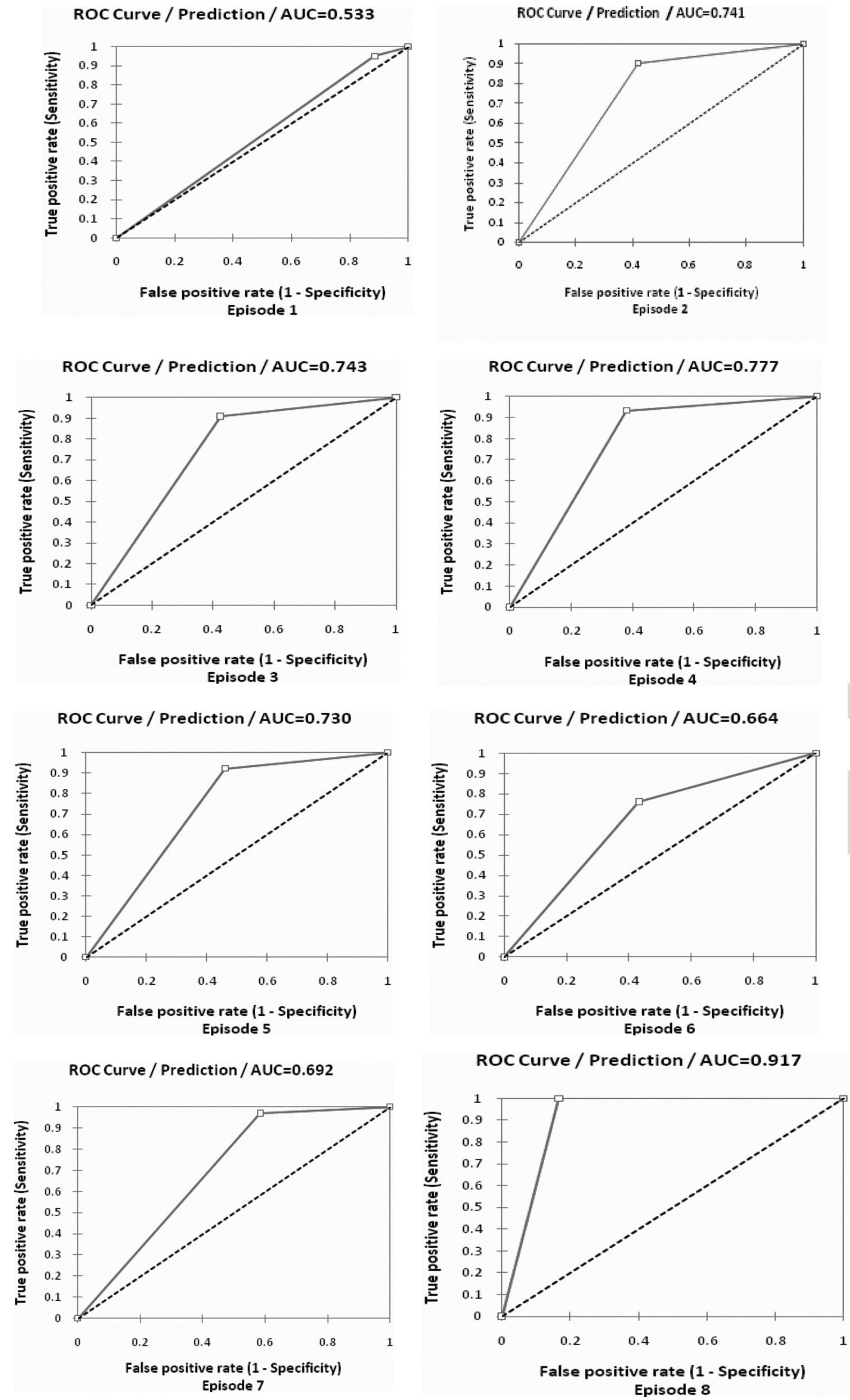

Fig. 5. ROC curve for Episodes one to eight for the machine learning algorithm.

The receiver operator characteristics (ROC) are summarized in Table IV (ACC: accuracy, NPV: negative predictive value, PPV: positive predictive value, SPC: specificity, SEN: sensitivity). The algorithm's final prediction of the screening attendance was based on a polling strategy that relies on the prediction confidence. The accuracy of the algorithm was around 68% for the first three episodes. Episode 4 had the maximum accuracy at 79%, closely followed by the fifth episode. The accuracies of the sixth and seventh episodes were lowest (57% and 51%, respectively). The NPV was the maximum at 51% for the fifth episode. The rest of the episodes had NPV values between 41% and 47%.

Episode 7 had the lowest NPV (30%). These lower NPVs were expected as the proportion of nonattendees was lesser in the test set (unbalanced). The PPVs for the fourth and fifth episodes were higher between 83% and 87%. The remaining episodes had values in the seventies range, except for the sixth episode where it was 64%. Specificity was highest for the seventh episode at 60%, but this may not be a true indicator as this episode had only 238 records in total. The next highest value was in the fifth episode at 49%. Episodes 1, 2, and 6 had values between 40% and 45%. Episodes 3 and 4 had lower values at 26% and 37%, respectively. The sensitivity was around 80% for the first four episodes, peaking at 85% for Episode 3. The higher the training set of records, the higher the sensitivity values. Since the previous algorithm (AI-ATT) had only four episodes, the averages for the first four episodes were used for comparing the JAABS and AI-ATT algorithms. The same set of attributes, when presented to commercial software (Clementine), generated improved results (see Table IV).

The first three episodes show an almost 10% increase in accuracy. Similarly, the later episodes (Episodes 4 and 5) when predicted by the JAABS–Clementine model, on average,do 6% better than the JAABS–Java algorithm, whereas Episodes 6 and 7 illustrated the maximum difference in accuracy (10–27%); this shows that the commercial software performed better even with a reduced training dataset. The NPV was lowest for the first episode, but was double when compared to AI-ATT and nearly 10% more than JAABS (Java). The NPV for the rest of the episodes (second to fifth) was around 73%. The remainder (sixth and seventh) were at 63% and 86%, respectively. The NPV is the metric that corresponds to the prediction of nonattendance and this was much better than that was achieved by the AI-ATT. Specificity is the next important measure and tests on Clementine showed promising results for all the episodes except for the first one.

The ROC curves for JAABS (Clementine) showed good prediction characteristics for all episodes except for Episode 1 (see Fig. 5). From the model's performance perspective, all these prediction characteristics were positive. The AI model proposed (JAABS—implemented in both Java and Clementine) was consistent and even outperformed the earlier model (AI-ATT) in many aspects. This could be attributed to the larger database and more complete attribute set and even the new predictor variable (screening distance) assisting in improving the algorithm's efficiency. The knowledge creation by applying AI (JAABS) is not only consistent, repeatable, and economical, but also ensures minimal human intervention. This is ideal for automating the whole process.

The proposed AI network (JAABS) for predicting screening nonattendance would be incorporated in a new breast screening software model that connects to the screening database to generate the screening batch. Based on the prediction, an automated message would be sent to the women's healthcare stakeholders (GPs, nurses, and other clinical specialists). These messages would be assimilated by the clinical system used by the stakeholders and would eventually flag the women as a nonattendee. When a woman's clinical record is opened, a flag/pop-up window would trigger opportunistic interventions that are aimed at educating the woman. This knowledge transfer would empower the woman to make an informed decision toward screening. This multistakeholder-based opportunistic intervention strategy would increase the overall breast screening attendance.

V. Conclusion

This paper discussed the details of how a machine learning-based prediction tool can be effectively applied to increase the breast cancer screening attendance. The need for a high degree of automation was highlighted to simplify the algorithm's adoption; such automation would also reduce overheads and make integration as seamless as possible [31]. From the model's performance perspective, all the prediction characteristics were positive. The machine learning-based AI model (JAABS—implemented in both Java and Clementine) proposed was consistent and even outperformed the earlier model (AI-ATT) in many aspects. The performance improvement could be attributed to the larger database, more complete attribute set and even the new predictor variable (screening distance). The knowledge creation by applying AI (JAABS) is not only reliable, repeatable, and economical, but also ensures minimal human intervention. There is still scope for improving the prediction efficiency and this can be achieved through better predictor attributes and/or improved machine learning techniques. The former would be difficult to achieve as the data source itself may not be available but the latter would be possible as better AI models, such as support vector machines, fuzzy logic, and genetic algorithms or a combination of these, would enable further investigation for increasing the efficiency.

Acknowledgment

The authors would like to thank J. Patnick CBE, Director, NHS Cancer Screening Programs (U.K.), for funding this research, Dr. M. Wallis, Consultant Radiologist, Cambridge Breast Unit team, and Margot Wheaton, Program Manager for the Warwickshire, Solihull and Coventry Breast Screening Service at Coventry and Warwickshire Hospital, for their excellent support and guidance throughout this research.

References

[1] American Cancer Society. (2010, Feb. 10). *Breast Cancer Facts & Figures 2009–2010* [Online]. Available: http://www.acsevents.org/downloads/STT/F861009_final%209-08-09.pdf.

[2] Cancer Research U.K. (2010, Feb. 10). *Breast Cancer—U.K. Mortality Statistics*. [Online]. Available: http://info.cancerresearchuk.org/cancerstats/types/breast/mortality/index.htm.

[3] NHS Breast Screening Programme—Cancer Screening Programmes Annual Review 2009. (2010, Feb. 10). [Online]. Available: http://www.cancerscreening.nhs.uk/breastscreen/publications/nhsbsp-annualreview2009.pdf.

[4] K. Turner, J. Wilson, and J. Gilbert, "Improving breast screening uptake: Persuading initial non-attenders to attend," *J. Med. Screening*, vol. 1,
Q2 pp. 199–202, 1994.

[5] A. Majeed, R. Given-Wilson, and E. Smith, "Impact of follow up letters on non-attenders for breast screening: A general practice based study," *J. Med. Screening*, vol. 4, pp. 19–20, 1997.

[6] J. P. Sin and A. S. Leger, "Interventions to increase breast screening uptake: Do they make any difference?," *J. Med. Screening*, vol. 6, no. 1, pp. 170–181, 1999.

[7] Canadian Cancer Society. (2006). *Canadian Researchers Find Common Breast Cancer Chemotherapy Regime Inferior at Preventing Disease Recurrence* [Online]. Available: http://www.cancer.ca/Canadawide/About%20us/Media%20centre/CW-Media%20releases/CW2006/Canadian%20Researchers%20Find%20Common%20Breast%20Cancer%20Chemotherapy%20Regime%20Inferior%20at%20Preventing%20Disease%20Recurrence.aspx?sc_lang=en.

[8] Canadian Cancer Society. (2008, Mar. 22). *Canadian Cancer Statistics 2008* [Online]. Available: http://www.cancer.ca/Canada-wide/About%20cancer/Cancer%20statistics/~/media/CCS/Canada%20wide/Files%20List/English%20files%20heading/pdf%20not%20in%20publications%20section/Canadian%20Cancer%20Society%20Statistics%20PDF%202008_614137951.ashx.

[9] A. Oikonomou, S. A. Amin, R. N. G. Naguib, A. Todman, and H. Al-Omishy, "Breast self examination training through the use of multimedia: A prototype multimedia application," *IEEE Eng. Med. Biol. Soc.*, vol. 2, no. 21, pp. 295–298, 2003.

[10] B. V. Marcela, "The system does work," *J. Am. College Radiol.*, vol. 1, no. 6, pp. 438–440, 2004.

[11] L. Wyld, "Mammographic Breast Screening in Elderly Women," in *Management of Breast Cancer in Older Women*, part 3, M. W. Reed and R. A. Audisio, Eds. London, U.K.: Springer, 2010, ch. 9, pp. 127–142.

[12] R. G. Blanks, S. M. Moss, C. E. McGahan, M. J. Quinn, and P. J. Babb, "Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990–1998: Comparison of observed with predicted mortality," *BMJ*, vol. 321, no. 7262, pp. 665–669, 2000.

[13] S. S. Epstein, *The Politics of Cancer*. New York: Doubleday, 1979, pp. 537.

[14] G. Burton, *Alternative Medicine*. Washington, DC: Future Medicine Publishing, 1997.

[15] Cancer Research U.K. (2007, Jul. 14). *Cancer Incidence—U.K. Statistics* [Online]. Available: http://info.cancerresearchuk.org/cancerstats/incidence/index.htm

[16] P. Forest, *Breast Cancer Screening—A Report to the Health Ministers of England, Scotland, Wales and Northern Ireland*. London, U.K.: HMSO, 1986.

[17] Medicine net (2010 Feb. 18). *Breast Cancer* [Online]. Available: http://www.medicinenet.com/breast_cancer/page3.htm

[18] I. Pirjo, L. Kauhava, I. Parvinen, H. Helenius, and P. Klemi, "Customer fee and participation in breast cancer screening," *The Lancet*, vol. 358, p. 1425, 2001.

[19] S. H. Woolf, "The 2009 Breast Cancer Screening Recommendations of the US Preventive Services Task Force," *JAMA*, vol. 303, no. 2, pp. 162–163, 2010.

[20] American Cancer Society Inc., (2010, Feb. 18) *Cancer Reference Information* [Online]. Available: http://www.cancer.org/docroot/CRI/CRI_2_5x.asp?dt=5

[21] D. P. Weller and C. Campbell, "Uptake in cancer screening programmes: A priority in cancer control," *Brit. J. Cancer*, vol. 101, pp. 55–59, 2009.

[22] Y. Zheng, "Breast cancer detection with gabor features from digital mammograms," *Algorithms*, vol. 3, pp. 44–62, 2010.

[23] K. W. Eilbert, K. Carroll, J. Peach, S. Khatoon, I. Basnett, and N. McCulloch, "Approaches to improving breast screening uptake: Evidence and experience from Tower Hamlets," *Brit. J. Cancer*, vol. 101, no. 2, pp. 64–67, 2009.

[24] D. Schopper and C. de Wolf, "How effective are breast cancer screening programmes by mammography? Review of the current evidence," *Eur. J. Cancer.*, vol. 45, no. 11, pp. 1916–1923, Jul. 2009.

[25] E. S. Cassandra, "Breast cancer screening: Cultural beliefs and diverse populations," *Health Soc. work*, vol. 31, no. 1, pp. 36–43, 2006.

[26] NHS Cancer Screening Programmes. (2007, Apr.) *Disclosure of Audit Results in Cancer Screening Advice on Best Practice* (Cancer Screening Series 3), J. Patnick, Ed. [Online]. Available: http://www.cancerscreening.nhs.uk/publications/cs3.pdf

[27] K. Okane. (2005, Apr. 20). *Mumps Language Bioinformatic Database Resources* [Online]. Available: http://bioinformatics.org/forums/forum.php?forum_id=1035

[28] V. Baskaran, R. K. Bali, R. N. G. Naguib, and H. Arochena, "A Knowledge Management approach to increase uptake in a breast screening programme," presented at the IEEE 2nd Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) Int. Conf., Philippines, Mar. 2005.

[29] S. Tarver, K. Cronin-Cowan, and M. E. Wheaton, "A pilot's life for us," *Breast Cancer Res.*, vol. 6, suppl. 1, p. 52, 2004.

[30] H. E. Arochena, "Modelling and prediction of parameters affecting attendance to the NHS breast cancer screening programme," Ph.D. dissertation, Dept. Comp. Sci., Coventry Univ., Coventry, U.K., 2003.

[31] C. Bankhead, S. H. Richards, T. Peters, D. Sharp, R. Hobbs, J. Brown, L. Roberts, C. Tydeman, V. Redman, J. Formby, S. Wilson, and J. Austoker, "Improving attendance for breast screening among recent non-attenders: A randomised controlled trial of two interventions in primary care," *J. Med. Screening*, vol. 8, no. 2, pp. 99–105, 2001.

Vikraman Baskaran is currently an Assistant Professor at the School of Information Technology Management of Ryerson University, Toronto, ON, Canada. His research interests include finding a viable application of the KM paradigm in healthcare application. His special interest in developing HL7 messaging and health informatics has provided opportunities in excelling in these fields. His current activities include KM, e-health, artificial intelligence, and healthcare informatics. Q3

He is a member of the HL7 U.K. and Canada. Q4

Aziz Guergachi (M'xx) is currently an Associate Professor at the Ted Rogers School of Information Technology Management of Ryerson University, Toronto, ON, Canada. Prior to becoming part of the Ryerson community, he was involved in the development of a large software system for trade promotion management and collaborative sales forecasting. His current research interests include advanced system modeling and machine learning with applications to business management and engineering systems.

He is the recipient of the New Opportunities Award of the Canada Foundation for Innovation and currently runs a research laboratory for advanced systems modeling. Q5

Rajeev K. Bali (SM'xx) is currently a Reader in Healthcare Knowledge Management at Coventry University, U.K. His main research interests include clinical and healthcare knowledge management (from both technical and organisational perspectives). He has published peer-reviewed journals and is the an author/editor of several textbooks on healthcare knowledge management.

He serves on various editorial boards and conference committees and is regularly invited to deliver presentations and speeches internationally. Q6

Raouf N. G. Naguib (SM'xx) is currently a Professor of Biomedical Computing and Head of BIOCORE, Coventry, U.K. Prior to this appointment, he was a Lecturer at Newcastle University, Newcastle Upon Tyne, U.K. He has published more than 240 journals and conference papers and reports in many aspects of biomedical and digital signal processing, image processing, artificial intelligence, and evolutionary computation in cancer research. Q7

He was awarded the Fulbright Cancer Fellowship in 1995–1996 when he carried out research at the University of Hawaii, Mãnoa, on the applications of artificial neural networks in breast cancer diagnosis and prognosis. He is a member of several national and international research committees and boards.

QUERIES

Q1. Author: Please check whether the edits made in the sentence “This large percentage of nonattendance not only. . .” retain your intended sense.

Q2. Author: Refs. [5], [6], [8], [12], [14], [15], [16], and [23] are not cited in the text. Please check and provide citations.

Q3. Author: Please provide the expansion of KM.

Q4. Author: Please provide the educational details of all the authors.

Q5. Author: Please provide the year in which Aziz Guergachi became “Member” of the IEEE.

Q6. Author: Please provide the year in which Rajeev K Bali became “Senior Member” of the IEEE.

Q7. Author: Please provide the year in which Raouf N. G. Naguib became “Senior Member” of the IEEE.