Discovering Potential Precursors of Mammography Abnormalities based on Textual Features, Frequencies, and Sequences

Robert M. Patton and Thomas E. Potok

Oak Ridge National Laboratory,
{pattonrm,potokte}@ornl.gov

Abstract. Diagnosing breast cancer from mammography reports is heavily dependant on the time sequences of the patient visits. In the work described, we take a longitudinal view of the text of a patient's mammogram reports to explore the existence of certain phrase patterns that indicate future abnormalities may exist for the patient. Our approach uses various text analysis techniques combined with Haar wavelets for the discovery and analysis of such precursor phrase patterns. We believe the results show significant promise for the early detection of breast cancer and other breast abnormalities.

1 Introduction

Most research involving mammography is performed on the image data, usually without regard to the corresponding textual reports. These reports are unstructured text, written by a human subject-matter expert, about the images of a patient. A set of reports for an individual patient forms a sequence of observations based on each patient visit. We believe that frequency analysis of phrase patterns can be analyzed to enhance the accuracy of detecting and forecasting breast anomalies. Our work focuses on a longitudinal view of the patients with respect to the mammogram reports. We believe a frequency analysis of the complete patient record will provide precursors to breast anomalies that currently cannot be detected. Consequently, this work seeks to explore the following questions:

- Do phrase patterns exist that act as precursors to future abnormalities in a patient?
- If so, how far in advance of the abnormality do they occur?

Answers to such questions may provide enhanced, automated detection as well as more efficient and effective care of patients. This paper describes initial work being performed to answer such questions. Section 2 discusses the background of the radiology reports being addressed by this work as well as the natural language processing that is needed. Section 3 discusses the analysis approach, while section 4 discusses results. Section 5 discusses conclusions.

2 Background

In this initial study, reports from 12,809 patients over a 5-year period are analyzed. There are 61,064 actual reports in this set, which include a number of reports that simply state that the patient canceled their appointment. For the work discussed here, we are particularly interested in studying the patients with multiple reports over time. Some of these patients have reports that predate the study and also have diagnostic screenings, indicating a potential health problem. Abnormal reports tend to have a richer, broader, and more variable vocabulary than normal reports. In addition, normal reports tend to have a higher number of "negation" phrases. These are phrases that begin with the word "no" such as the phrase "no findings suggestive of malignancy"[3]. Another challenge in analyzing the natural language of these reports is that there are multiple ways of conveying the same meaning. Phrases such as "no strongly suspicious masses" and "no new suspicious mass lesions" both mean that nothing cancerous was observed. To account for this variability in the language, we used the natural language processing technique known as skip bigrams, or s-grams. S-grams are word pairs in their respective sentence order that allow for arbitrary gaps between the words[6]. For example, the s-gram for the previous phrase examples is the words *no* and *suspicious*. As discussed in [4], s-grams are an effective technique in determining normal and abnormal reports. The next step is to analyze patient records to determine if abnormal report occurrences can be forecasted.

3 Approach

The objective of this work is to identify whether certain phrase patterns exist within patient reports that act as precursors to future abnormalities. To explore this, we analyzed the patient records that contain a higher number of reports. This narrowed the patient data set to 667 patients who had between 12 and 16 reports each. Of these patients, the ones of most interest for this work are those with discernable patterns in their medical reports. For example, a patient with cancer may have had an early report that mentions something unusual or suspicious, followed by years of normal reports, before the cancer appeared. The early report may be a precursor for the cancer. To identify these patients, each report in each patient record is analyzed to count the number of normal and abnormal s-grams as described in [4]. This provided a temporal sequence of normal and abnormal s-gram counts for each patient record. The following table shows an example patient record where s-grams were counted for each report.

In the example shown in Table 1, there is some abnormality that is mentioned early in the record (May 24, 1984) and then the record contains multiple abnormal s-grams toward the end of the record (beginning on Dec 7, 1991). The normal and abnormal s-gram counts form a temporal sequence for each patient. Our goal is to be able to compare patients based on these sequences. To find patients with similar patterns in the set of 667, we use a discrete wavelet transform (DWT) of the temporal sequence of abnormal s-gram counts [1][5]. A wavelet

Table 1. Example record of normal and abnormal s-gram counts for patient A

Mammogram Date	Normal S-grams	Abnormal S-grams
May 20, 1981	1	0
May 24, 1984	3	1
June 3, 1985	3	0
March 9, 1988	1	0
July 12, 1989	4	0
Dec 5, 1990	3	0
Dec 7, 1991	1	4
March 11, 1992	0	4
March 11, 1992	0	4
March 22, 1992	0	1
March 22, 1992	0	1
March 23, 1992	0	0
Nov 9, 1992	0	0

transform is a mathematical function that is used to split a function into separate scale components, thus providing a multi-resolution analysis. The wavelet transform is analogous to a prism that breaks light into its various spectral colors. They are widely used in time-series analysis, as well as in other domains such as image processing. A critical feature of the DWT is that it will not only identify the frequencies that constitute a temporal sequence, but also the location in time in which those frequencies occur. It is this feature of the DWT that is exploited in this work, as our objective is to find phrase patterns that occur prior to other phrase patterns. In addition, a DWT provides the ability to find similar temporal patterns, allowing for the flexibility of matching patterns despite amplitude and time shifts. Previous work has shown wavelets to be effective in performing similarity searches of time series [2]. However, the work described here utilizes a rule-based approached to finding similar temporal patterns using DWT that does not rely on the use of thresholds. This enables a wider range of temporal patterns to be found that contain the basic temporal characteristics of interest. Each patient record consisted of 16 or fewer reports. For records with less than 16 reports, the temporal sequences were padded with zeros until there were 16 elements. Next, for each patient record, the temporal sequence of abnormal s-gram counts were transformed using a Haar wavelet [1]. For example, the transform for patient *A* (of Table 1) is shown in the following table. After each

Table 2. Haar wavelet transform of abnormal s-gram sequence for patient A

1st coefficient 0.9375								
Band 0	0.1875							
Band 1	-0.875	0.75						
Band 2	0.25	-2	1	0				
Band 3	-0.5	0	0	0	1.5	0.5	0	0

of the 667 patient records is transformed via a Haar wavelet, the next step is to begin looking for the patterns of interest, early abnormality and late anomaly. First, resolution *1* of each patient is examined. Specifically, the first coefficient

of resolution *1* should be less than 0 while the second coefficient of resolution *1* should be greater than 0. This particular pattern identifies those patients with an increasing amplitude change in their s-gram counts toward the end of the records (rather than at the beginning of their records), which suggests that diagnostic screening was performed near the end of the patient's record. Second, if the pattern for resolution *1* exists, then resolution *2* of each patient is examined. Specifically, either the first or second coefficient (or both) of resolution *2* should be less than 0 while the third and fourth coefficients should both be greater than 0. This particular pattern identifies those patients who have a short duration of abnormal s-gram counts early in the records, which suggests that some unusual feature about the patient was mentioned early in their record. For higher resolution, resolution *3* could be used instead of resolution *2*. In that case, the first four coefficients would be checked for negative values, while the last four coefficients would need to be positive. Patient records that match these patterns in the Haar DWT are then selected. This reduced the data set to 123 patient records, which is approximately 1% of the original data set. For these selected patient records, all s-grams were extracted from the first report in which the abnormal s-gram count was at least 1 but less than or equal to the normal s-gram count. This represents a normal report where some potential abnormality was mentioned. Next, the time elapsed was computed between this first report and the next report where the abnormal s-gram count was higher than the normal s-gram count. This second report represents an abnormality that was detected and a diagnostic screening was requested. From the example data shown of patient A in Table 1, the first report would be the one dated May 24, 1984 and the next report would be the one dated December 7, 1991. All s-grams from the report dated May 24, 1984 are extracted and considered as potential precursor patterns. Finally, the frequency of each extracted s-gram was computed along with the corresponding average elapsed time. The results of this approach applied to the 123 selected patients are shown in the following tables and figures.

4 Results

Table 3 shows the top three precursor s-grams that were observed. In reviewing this table, there is no single definitive precursor s-gram. However, the top three s-grams have approximately a fifty percent occurrence as a precursor. This means that, of the 123 selected patients, if one of those s-grams were mentioned in the patient's record, then there is a fifty percent chance that the patient will have a diagnostic screening (i.e., an abnormality will be seen that requires additional testing) at some point in the future. While this percentage is equivalent to random selection, in comparison to the other s-grams found, these s-grams show promise as potential precursors and demonstrate a capability far beyond the current state of the practice, which is dependent entirely on manual analysis. Table 4 shows the average elapsed time in units of days for each of the s-grams shown in Table 3. What is very encouraging in these results is that the first and third s-grams provide approximately a three to five year lead-time. This

Table 3. Top three precursor s-grams

S-gram	Occurrences as Precursor	Occurrences in Selected Patients	% Occurrence as Precursor
lymph & node	39	71	54.93
cm & density	12	24	50.00
nodular & density	51	104	49.04

provides a very early warning indication of a future abnormality. The drawback, however, is that the skewness and kurtosis values for these s-grams indicate significant variability in this window. The reason for this is that these terms are general and vague in their meaning, but still provide some level of indication that the radiologist sees a feature of concern. In contrast, the second s-gram (*cm & density*) provides a much more specific window with an average of just over one year with very high positive skewness and kurtosis values. The reason for this is that this s-gram represents phrase patterns that are very specific about a particular feature that was observed in the patient. An example phrase that this s-gram would represent is "*2.5 cm area of asymmetric density*". Such specificity by the radiologist suggests that the radiologist is very focused on this feature and is likely to be concerned enough to request additional diagnostic screenings. Consequently, the average time elapsed for this s-gram is much shorter and has less variability. The data in Table 5 shows the usage frequency of the s-grams

Table 4. Precursor lead-time

S-gram	Average Time Elapsed (years)	Std Dev (years)	Skewness / Kurtosis
lymph & node	4.2	2.9	0.01 / -1.38
cm & density	1.1	2.2	2.63 / 6.91
nodular & density	2.9	2.9	0.68 / -0.64

shown in Table 3. In document text analysis, terms and phrases that are commonly used in a document set are not considered useful in characterizing the content of a particular document. However, if a term or phrase is not commonly used in a document set and a particular document has a high frequency of that term or phrase, then it is considered significant to that document. In a similar manner, the frequency of s-grams in Table 3 were computed over all of the patients (12,809 patients) and over the patients that were selected for analysis (123 patients). These frequencies, as well as the corresponding percent increase, are shown in Table 5. As can be seen, most of the s-grams have percent increases well over 100%. This is encouraging as it shows that these s-grams are highly related to patients with abnormalities. If the percent increases had been much below 100%, then this would indicate that these s-grams are very common, and consequently, the value as a precursor would be diminished. However, the percent increases and corresponding percent occurrence in selected patients shown in the table suggest that these s-grams have high potential as precursors.

Table 5. Comparison of s-gram usage frequency

S-gram	% Occurrence in All Patients	% Occurrence in Selected Patients	% Increase in Occurrence
lymph & node	25.17	57.72	129.34
cm & density	5.50	19.51	254.51
nodular & density	31.39	84.55	169.35

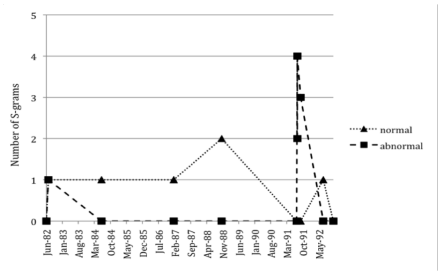

Fig. 1. Example patient record with "lymph" & "node" as a precursor s-gram

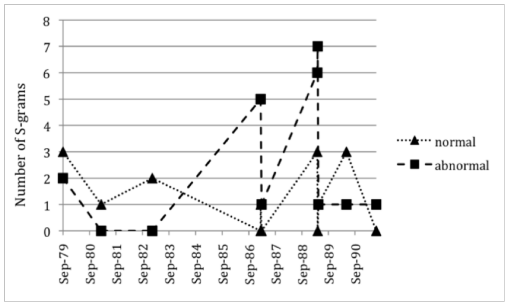

Fig. 2. Example patient record with "cm" & "density" as a precursor s-gram

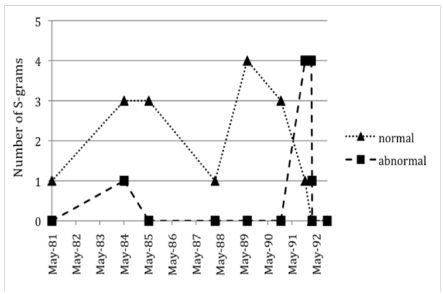

Fig. 3. Example patient record with "nodular" & "density" as a precursor s-gram

The figures show various patient records that were found using the approach described here. Figure 1 shows the normal and abnormal s-gram counts of a patient record found by this approach where "*lymph & node*" was a precursor s-gram. In one of the first reports in this record, a radiologist made particular note of specific lymph nodes in this patient. This patient was ultimately diagnosed with grade 1 infiltrating ductal carcinoma (i.e., breast cancer) with tubular differentiation. Figure 2 shows the normal and abnormal s-gram counts of another patient record found by this approach where "*cm & density*" was a precursor s-gram. In the first report of this record, the radiologist states "There is a less than 1 cm area of focal increased density seen only on the left craniocaudal view in the lateral aspect of the left breast." This patient was ultimately diagnosed with "mild fibrocystic disease with radial scar and focal florid sclerosing adenosis" in the right breast. Figure 3 shows the normal and abnormal s-gram counts of another patient record found by this approach where "*nodular & density*" was a precursor s-gram. In one of the first reports, the radiologist states "There is prominent nodular density posteriorly and inferiorly in both breasts on the mediolateral oblique views, left more than right." This patient is ultimately diagnosed with a simple cyst. In that report, the radiologist states "Ultrasound directed to the inferocentral left breast 6 o'clock position demonstrates a 1-cm round, simple cyst." In each of these examples, it should be noted that the precursor s-gram does not necessarily provide specific information concerning the abnormality that is ultimately diagnosed. In the first two examples, the s-grams are not related to the ultimate diagnosis. In the third example, the precursor s-gram is related, but it cannot be conclusively determined that it is, in fact, the exact same abnormality that is ultimately diagnosed. However, what the precursor s-gram does provide is an early warning indication that the radiologist noted some feature about the patient that seemed unusual, or was noteworthy. The approach described here seeks to leverage that information, even if it does not ultimately relate to the final diagnosis.

5 Conclusions

The initial objective of this work was to answer the following questions:

- Do certain phrase patterns exist that act as precursors to future abnormalities in a patient?
- If so, how far in advance of the abnormality do they occur?

As can be seen in the results, phrase patterns do exist that act as precursors. In addition, these precursors also hold the potential of providing lead times measured in years. This is potentially very significant, although additional work is needed to investigate this possibility. In this work, there are several other positive outcomes. First, in the approach developed, abnormal reports are identified based on s-grams related to diagnostic screenings, not based on specific types of abnormalities. Consequently, the precursor s-grams provide a general warning indication. Any form of early warning detection will provide various levels of

specificity. This preliminary work provides the initial level of warning. Second, the results show that the precursor s-grams are used much more frequently in patients with abnormalities in comparison to the entire set of patients. This is significant in that it provides confidence that these precursor s-grams are, in fact, related to abnormalities.

6 Acknowledgements

Our thanks to Robert M. Nishikawa, Ph.D., Department of Radiology, University of Chicago for providing the large dataset of unstructured mammography reports, from which the test subset was chosen.

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. De-AC05-00OR22725.

This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

1. Burrus, C.S., R.A. Gopinath, and H. Guo, Introduction to Wavelets and Wavelet Transforms, A Primer. Prentice Hall, 1997.
2. Chan, F.K.-P., A.W.-C. Fu, and C. Yu, "Haar wavelets for efficient similarity search of time-series: with and without time warping," IEEE Trans. On Knowledge and Data Engineering, vol. 15, no. 3, May-June 2003.
3. Patton, R.M., B.G., Beckerman, and T.E. Potok, 2008. "Analysis of mammography reports using maximum variation sampling." Proceedings of the 4th GECCO Workshop on Medical Applications of Genetic and Evolutionary Computation (MedGEC), Atlanta, USA, July 2008. ACM Press, New York, NY.
4. Patton, R.M., B.G. Beckerman, J.N. Treadwell, and T.E. Potok, 2009. "A Genetic Algorithm for Learning Significant Phrase Patterns in Radiology Reports." Proceedings of the 5th GECCO Workshop on Medical Applications of Genetic and Evolutionary Computation (MedGEC), Montreal, Canada, July 2009. ACM Press, New York, NY.
5. Percival, D.B., and A.T. Walden, Wavelet methods for time series analysis, Cambridge University Press, 2000.
6. Pirkola, A., H. Keskustalo, E. Leppanen, A. Kansala, and K. Jarvelin, 2002. "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants." Information Research, 7(2) [Available at http://InformationR.net/ir/7-2/paper126.html]