Analysis and Classification of Mammography Reports Using Maximum Variation Sampling

Robert M. Patton, Barbara G. Beckerman, Thomas E. Potok

1. INTRODUCTION

Currently, no automated means of detecting abnormal mammograms exist. While knowledge discovery capabilities through data mining and data analytics tools are widespread in many industries, the healthcare industry as a whole lags far behind. Providers are only just beginning to recognize the value of data mining as a tool to analyze patient care and clinical outcomes [8]. The research conducted by the authors investigates the use of genetic algorithms for classification of unstructured mammography reports, which can be later correlated to the images for extraction and testing.

In mammography, much effort has been expended to characterize findings in the radiology reports. Various computer-assisted technologies have been developed to assist radiologists in detecting cancer; however, the algorithms still lack high degrees of sensitivity and specificity, and must undergo machine learning against a training set with known pathologies in order to further refine the algorithms with higher validity of truth. In a large database of reports and corresponding images, automated tools are needed just to determine which data to include in the training set. Validation of these data is another issue. Radiologists disagree with each other over the characteristics and features of what constitutes a normal mammogram and the terminology to use in the associated radiology report. Abnormal reports follow the lexicon established by the American College of radiology Breast Imaging Reporting and Data System (Bi-RADS) [2], but even within these reports, there is a high degree of text variability and interpretation of semantics. The focus has been on classifying abnormal or suspicious reports, but even this process needs further layers of clustering and gradation, so that individual lesions can be more effectively classified [29]. The tools that are needed will not only help further identify problem areas but also support risk assessment and other knowledge discovery applications.

The knowledge to be gained by extracting and integrating meaningful information from radiology reports will have a far-reaching benefit, in terms of the refinement of the classifications of various findings within the reports. This will support validation, training and optimization of these and other machine learning and computer-aided diagnosis algorithms to work both in this environment and with other medical and imaging modalities. In the near-term, the objective of this work is to accurately identify abnormal radiology reports amid a massive collection of reports. The challenge in achieving this objective lies in the use of natural language to describe the patient's condition. The premise of this work is that abnormal radiology reports consist of words and phrases that are statistically rare or unusual. If this is true, then it is expected that abnormal reports will be significantly dissimilar in comparison to normal radiology reports.

To achieve this objective, our approach employs maximum variation sampling (MVS), which is implemented as an adaptive sampling approach[16][32][33]. Maximum variation sampling seeks to identify a particular sample of data that will represent the diverse data points in a data set. Adaptive sampling continues to draw samples from the population based on previous samples until some criteria have been met. Previous results from using MVS indicated that an ideal sample could be found very quickly using this approach [17][18].

2. BACKGROUND

Mammography is the procedure of using low-dose X-rays to examine the human breast for the purposes of identifying breast cancer or other abnormalities. Currently, for each patient that undergoes a mammogram, there is at least one X-ray image and one textual report written by a radiologist. In the report, the radiologist describes the features or structures that they see or do not see in the image. If an abnormality or suspicious area is found, the patient may undergo a diagnostic mammogram or biopsy, which results in additional images and reports in the patient's record. Essentially, these reports are meta-data about the corresponding image that is written by a human subject matter expert. In order to effectively train a computer-assisted detection (CAD) system, these reports could be mined and used as supplemental meta-data. Unfortunately, little work has been done to utilize and maximize the knowledge potential that exists in these reports.

There are several challenges in utilizing these reports. First, the reports vary in length. Some radiologists use more words than others when describing the same features. For example, in patients that do not exhibit any suspicious features, there are some reports that very simply state that there are no suspicious features. However, for the exact same patient with no suspicious features in a different year, a different radiologist will provide a much more lengthy report that describes all of the suspicious features that did not exist.

To provide a better perspective of the challenge of mining these reports, consider the following question. Given a database of these reports, how does one classify those reports that represent abnormalities in the patient? In mammography, most patient reports will represent "normal" conditions in the patient. Consequently, the reports with "abnormal" conditions are rare (defining the difference between what is "normal" and "abnormal" is beyond the scope of this work). Performing a cluster of these reports, most of the normal reports would cluster together while the abnormal reports would not form a cluster. This is because "abnormal" conditions tend to be very unique and very specific to a patient while "normal" conditions are much more generic and broad. Even if clustering provided value, clustering a very large database of these reports is exceptionally computationally expensive. Categorizing would be faster, however, the challenge remains of determining the appropriate categories, and even then, the abnormal reports may not categorize correctly.

Another challenge to utilizing these reports lies in the language that is used in mammograms. Abnormal reports tend to have a richer 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 in the phrase "no findings suggestive of malignancy." Consider the phrases shown in Table 1 and Table 2. These are the

negation phrases that generally occur in normal reports and the ones shown here are samples of the variations that have been found. In the set of reports used for this work, there were at least 286 variations of phrases for Table 1 and 1,231 variations of phrases for Table 2.

Table 1. Example phrases using "no" and "malignancy"

no malignancy
no mammographic features malignancy
no mammographic features suggestive of malignancy
no findings suggestive of malignancy
no significant radiographic features of malignancy
no radiographic findings suggestive of malignancy
no radiographic change suggestive of malignancy
no specific radiograpic features of malignancy
no mamographic evidence of malignancy

Table 2. Example phrases using "no" and "suspicious"

no mammographic finding suspicious
no strongly suspicious forms
no strongly suspicious features
no strongly suspicious masses
no radiographically suspicious masses
no developing suspicious clustered microcalcifications
no finding strongly suspicious
no new suspicious mass lesions
no suspicious linear branching forms

Consider the phrase shown in Table 3 and Table 4. These phrases tend to occur in abnormal reports (but may also occur in normal reports) and the ones shown here are samples of the variations that have been found. In the set of reports used for this work, there were at least 52 variations of phrases for Table 3 and 691 variations of phrases for Table 4.

Table 3. Example phrases using "appearance" and "tissue"

appearance suggesting radiating strands of tissue
appearance suggestive of accessory breast tissue
appearance of normal glandular tissue
appearance of asymmetric fibroglandular tissue
appearance of fibroglandular tissue
appearance of glandular tissue
appearance of normal fibroglandular tissue
appearance of soft tissue densities bilaterally

Table 4. Example phrases using "additional" and "views"

additional views obtained today demonstrate variation
additional compression views
additional set of bilateral cc views
additional lateral views

additional mediolateral oblique views
additional mammographic views
additional bilateral craniocaudal views
additional bilateral lateral medial views

Considering the language variations shown previously, the task of classifying those reports that represent abnormalities is daunting. The variations of terms and syntax create a combinatorial explosion while, semantically, these combinations tend to mean the same thing.

3. RELATED WORKS

There has been considerable work in a variety of areas in the text analysis community and a wide array of problems with processing and analyzing text data. Some of these areas of text analysis include retrieval, categorization, clustering, syntactic and semantic analysis, duplicate detection and removal, and information extraction to name a few [4][23][24][28][34]. These areas range from analyzing entire datasets to analyzing a single document. In general, as the size of the dataset increases, many of these approaches begin performing poorly, or the value of their results begins to diminish. For example, clustering usually requires comparing every document with every other document. Obviously, as the dataset size increases, performance will noticeably suffer. However, with categorization, the performance may not suffer considerably, but the quality of the results will be diminished if a sufficient number of categories are not identified or if the categories are not clearly or accurately identified [27].

Further improvement in information retrieval techniques requires the continued development of algorithms whose basis lies in semantic extraction and representation. Information retrieval (IR) research began with simple representations of documents and the terms that they contained [26]. This research progressed into syntactic analysis such as co-occurrence, N-grams, part of speech analysis, and context-free grammars. Recently, IR research has continued to move toward a basis in semantics. Many of these approaches involve the use of ontologies, conceptual graphs, and language models such as described in [6][7][11][12][26]. Unfortunately, many of these approaches are either unable to scale, require significant effort on the part of subject matter experts, or do not handle domain specific data robustly. The work described here differs from these approaches in that it leverages computationally efficient, unsupervised learning of domain specific data in order to more effectively retrieve information. As a result, extensive ontologies are not needed, or extensive effort on the part of a subject matter expert.

In [1], an unsupervised approach to identifying cue phrases is discussed. Cue phrases are formulaic patterns of phrases that have similar semantics but vary in syntactical and lexical ways. In [1], the authors use a lexical bootstrapping algorithm that relies on the use of “seed” phrases. While our work is addressing nearly the same problem, our work differs in that no seed phrases are needed, and that s-grams found for cue phrases using our approach are split into two classes.

Other work is being done in the medical environment to use automated software tools to extract knowledge from unstructured radiology reports [5]. Preliminary findings demonstrate that automated tools can be used to validate clinically important findings and recommendations for subsequent action from unstructured radiology reports. Commercially available software is also being tested to automate a method for the categorization of narrative text radiology reports, in this case dealing with the spine and extremities [31].

4. MAXIMUM VARIATION SAMPLING

The objective of this work is to accurately identify abnormal radiology reports amid a massive collection of reports. As discussed earlier, abnormal reports have wider variation in their language than normal reports. Consequently, what is needed is to sample the most diverse reports and identify the common language that is unique to those reports. This common language will then provide the basis for classification.

Sampling can be divided into two main categories: probability-based and nonprobability-based. Probability-based sampling is based on probability theory and the random selection of data points from the dataset. Nonprobability-based sampling is based on purposeful selection, rather than random selection. The advantage of this form of sampling is that it allows the analyst to look at data that may not otherwise be visible via the random selection process. In the domain of mammography reports, random selection would not easily find abnormal reports, as they constitute a very small portion of all reports.

Within nonprobability-based sampling, there are several categories of sampling [16], one of which is maximum variation sampling (MVS) [16]. This particular sampling method seeks to identify a particular sample of data that will represent the diverse data points in a data set. According to Patton [16], "This strategy for purposeful sampling aims at capturing and describing the central themes or principle outcomes that cut across a great deal of [data] variation." In a large text corpus, this form of sampling provides the ability to quickly characterize the different topics, or "threads" of information that are available.

A genetic algorithm (GA) was developed to implement the maximum variation sampling technique. Genetic algorithms (GA) are nature-inspired algorithms that mimic the natural selection process [10]. The natural selection process is generically defined as survival of the fittest (i.e., only the most fit individuals for a given environment survive and reproduce offspring). During this process, the offspring are created from the best individuals; therefore, the population should continue to improve over several generations. It has been shown that canonical genetic algorithms converge to an optimal solution if the best individual remains in the population [25].

It is well known that a genetic algorithm performs very well for large search spaces and is easily scalable to the size of the data set. In addition, GAs are also particularly suited for parallelization [13][15][30]. To better understand the need for scalability and the size of the search space in this problem domain, consider a set of 10,000 radiology reports. Now, suppose an analyst needs to reduce this data set to 200 representative reports (only 2% of the entire data set). In that case, there are approximately 1.7×10^{424} different

combinations of reports that could be used to create a single sample. Clearly, a brute force approach is unacceptable. In addition, many of the combinations would consist of duplicate data that would lower the quality of the result for the analysts. Ultimately, an intelligent and scalable approach such as a genetic algorithm is needed. As demonstrated by Mutalik [14], a parallel genetic algorithm is well suited to a combinatorial optimization problem.

Before applying a GA to the analysis of radiology reports, the reports must be prepared using standard information retrieval techniques. First, reports are processed by removing stop words and applying the Porter stemming algorithm [9][20][21]. Once this has been done, the articles are then transformed into a vector-space model (VSM) [22][26]. In a VSM, a frequency vector of word occurrences within each report can represent each report. Once vector-space models have been created, the GA can then be applied.

Two of the most critical components of implementing a GA are the encoding of the problem domain into the GA population and the fitness function to be used for evaluating individuals in the population. To encode the data for this particular problem domain, each individual in the population represents one sample of size *N*. Each individual consists of *N* genes where each gene represents one radiology report (each report is given a unique numeric identifier) in the sample. For example, if the sample size were 10, each individual would represent one possible sample and consist of 10 genes that represent 10 different reports. This representation is shown in the following figure.

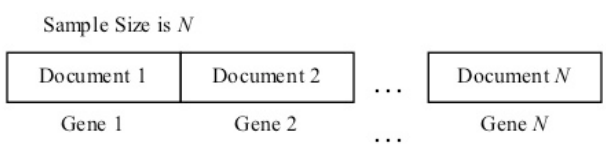

Figure 1. Genetic representation of each individual

The fitness function evaluates each individual according to some predefined set of constraints or goals. In this particular application, the goal for the fitness function was to achieve a sample that represents the maximum variation of the data set without applying clustering techniques or without prior knowledge of the population categories. To measure the variation (or diversity) of our samples, the summation of the similarity between the vector-space models of each document (or gene) in the sample is calculated as shown in the following equation.

$$Fitness(i) = \sum_{j=0}^{N} \sum_{k=j+1}^{N} \alpha_j + \beta_k + Similarity\,(Gene(i,j), Gene(i,k))$$

Equation 1. MVS-GA Fitness Function

In Equation 1, the Similarity function calculates the distance between the vector space models of gene *j* and *k* of the individual *i*. This distance value ranges between 0 and 1 with 1 indicating that the two reports are identical and 0 indicating that they are completely different in terms of the words used in that report. Therefore, in order to find a sample with the maximum variation, Equation 1 must be minimized (i.e., lower fitness

values are better). In this fitness function, there will be $(N^2 - N) / 2$ comparisons for each sample to be evaluated.

In an effort to effectively characterize the phrase patterns of the mammography reports, it is necessary to examine reports that are longer in length, so that more language can be examined for patterns. The data set for this work contains numerous reports that simply state that the patient canceled their appointment. These reports are very short in length and are exceptionally distinct from all other reports (similarity values approaching zero). In addition, abnormal reports tend to be longer in length than normal reports since the radiologist is describing the abnormalities in more detail. Consequently, the fitness function of the MVS-GA incorporates penalty functions as shown in equations 2 and 3.

$$\alpha_j = e^{-\left(\frac{\|j\|}{100}\right)}$$

Equation 2. Penalty factor for document *j*

$$\beta_k = e^{-\left(\frac{\|k\|}{100}\right)}$$

Equation 3. Penalty factor for document *k*

In the penalty equations, shorter documents receive higher penalties while longer documents receive much lower penalties. The penalty functions also return values that are between 0 and 1, inclusive. As a result of the penalty functions, the cancellation reports will receive the highest fitness values, while lengthy, abnormal reports will receive the lowest fitness values.

To create children from a given population, genetic operators such as selection, crossover, and mutation are applied to the individuals. For each generation, an average fitness value is calculated for the population. Individuals with fitness values that are above this average are selected as parents, while the other individuals are discarded. This can be a very aggressive selection process if there are extremely fit individuals that are far above the average. Once parents are selected, crossover and mutation operators are applied to the parents to create children. The crossover and mutation operators are 1-point operators [10]. After the MVS-GA is executed, the end result is a best sample of mammography reports that are as diverse from each other as possible.

In addition to finding a sample of the most diverse reports, the MVS-GA was also enhanced to extract the common skip bigrams (s-grams) from the reports. S-grams are word pairs in their respective sentence order that allow for arbitrary gaps between the words [1][3][19]. The s-grams for Table 1 in Section 2 are the words "no" and "malignancy." This s-gram uniquely identifies a particular semantic in the language of mammography reports and enables the identification of all possible variations of such phrases. Higher-level patterns may then be formed from these s-grams. For example, the s-grams for Table 1 and Table 2 in Section 2 both imply that there are no abnormalities seen in the patient.

Once the best sample is achieved by the MVS-GA, then phrases are extracted from each document in the sample. For each phrase in the document, s-grams are extracted. Next, the s-grams are counted across the sample of documents. S-grams that are common across the sample will have higher frequency counts while s-grams with a frequency of 1 uniquely identify a particular document in the sample. For this work, only those s-grams that are the most frequent in the best sample found are considered valuable. It is these s-grams that have the ability to uniquely classify abnormal documents from a large set.

The primary intent of the GA is to converge toward an optimal solution. However, very little GA research, if any, has been performed that leverages knowledge gained from the individuals that failed to be selected and reproduce. In a typical GA, individuals that are not selected for reproduction are simply discarded. For the MVS-GA, this is different. In addition to extracting s-grams from successful individuals, the MVS-GA has also been augmented to store the most frequent s-grams of the failed individuals. This will enable answering questions such as what characteristic phrases make failed individuals inferior to successful individuals. After each generation, s-grams and their frequencies from each failed individual are extracted from each individual and stored in memory. After the MVS-GA has completed, the memory now contains the most frequent s-grams that caused individuals to fail in the GA. Individuals that fail in the MVS-GA tend to contain a high number of normal reports. Successful individuals tend to contain a high number of abnormal reports. The end result is that the MVS-GA learns the most frequent s-grams for both abnormal and normal classes of reports.

5. DATA

In this work, unstructured mammography reports were used. Each report generally consists of two sections. The first section describes what features the radiologist does or does not observe in the image. The second section provides the radiologist's formal medical opinion as to whether or not there are suspicious features that may suggest malignancy (i.e., or the possibility that the patient has cancer).

These reports represented 12,809 patients studied over a 5-year period from 1988 to 1993. There are 61,064 actual reports in this set, which include a number of reports that simply state that the patient canceled their appointment. Table 1 and Figure 1 shows the general statistics and distribution of the number of reports per patient.

Table 1. Statistics of Number of Reports per Patient

Minimum	1
Maximum	32
Average	4.77
Std Dev	3.57
Skewness	1.27
Kurtosis	1.76

According to Table 1, the positive skewness indicates that there are many patients with more reports than the average. Since the study was five years in duration, the vast majority of the patients will have either 1 report every year or 1 report every 2 years.

This explains the average of 4.77 and kurtosis of 1.76. For some of the patients, abnormalities were identified that required additional diagnostic screenings. In addition, some patients have reports that predate the beginning of the study. In the extreme case of 32 reports, a patient record contained reports that predated the study by nearly 10 years and the patient also had breast cancer in the right breast and an abnormality in the left breast that was later determined to be benign. Consequently, these patients with the extra reports explain the positive skewness.

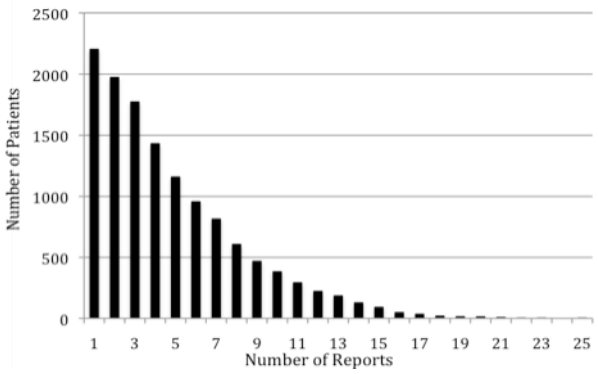

Figure 1. Distribution of the Number of Reports per Patient

Furthermore, of this large data set, a human expert manually classified 100 reports as being normal and 100 reports as being abnormal. From the normal set of reports, a random sample of 90 reports was selected. From the abnormal set, a random sample of 10 reports was selected. The two samples were then merged to create a third set of 100 reports. This third set was used to initially test the MVS-GA. If the premise that abnormal radiology reports consist of words and phrases that are statistically rare or unusual, then the expected result of the MVS-GA will be a sample of reports consisting predominantly of abnormal reports.

6. TESTS

Two sets of tests were performed. For the first set of tests, the test data consisted of just 100 reports from the human classified group, where 90 of these reports were classified as normal and 10 were classified as abnormal. The objective of the first set of tests was to determine if, in fact, the MVS-GA could identify abnormal reports amid a significantly larger number of normal reports. After the first tests were performed, the MVS-GA was enhanced as described early to extract s-grams from the successful and failed individuals in the population. The objective of the second set of tests was to learn the key s-grams that could be used to classify normal and abnormal reports.

For the initial tests, thirty runs of the MVS-GA were performed. The population size was defined as 2,000 and the number of generations was set to 250. The crossover rate was set to 0.7 and the mutation rate was set to 0.03. The sample size was set to 15 and the data set size was 100 radiology reports (90 normal and 10 abnormal). In this case, there are approximately 2.53×10^{17} different combinations of reports that could be used to create a single sample.

For the second set of tests, the MVS-GA was enhanced to extract s-grams from both successful and failed individuals in the population. The population size was defined as 2,000 and the number of generations was set to 2,500. The crossover rate was set to 0.7 and the mutation rate was set to 0.03. The sample size was set to 100 and the data set size was the entire set of 61,064 reports. In this case, there are approximately 3.74×10^{320} different combinations of reports that could be used to create a single sample.

7. RESULTS & DISCUSSION

For the initial tests, the GA consistently found 8 out of 10 abnormal reports. The remainder of the sample consisted of 7 normal reports. Upon further analysis of the 10 abnormal documents, it was found that 4 of the reports were very similar to each other, while the other 6 were very distinct from each other. Consequently, 2 of the 10 abnormal reports were consistently absent from the final sample.

Upon further analysis of the normal documents that were included in the final sample, it was determined that several of the reports represented "boundary" cases. These were reports that, while considered normal by a human expert, represented situations where a patient had either already undergone a lumpectomy or had a family history of breast cancer and showed high potential for breast cancer. Other normal reports that were in the final sample consisted of patients that needed further examination and therefore underwent spot magnification for further confirmation. Another report represented a patient where the radiologist had difficulty in determining a nodule in the image and suggested that it was a "small deformable cyst." Overall, these preliminary results from the GA showed encouraging performance to find both abnormal reports and potentially unusual normal reports without prior categorization or a predefined vocabulary of terms to search.

Additional analysis of the final sample revealed another characteristic of the reports. For each report, word phrases unique to that specific report were extracted. In this case, unique word phrases are those phrases that only appear in one report in the sample. As shown in Table 1, normal reports tended to have fewer unique word phrases as compared to abnormal reports. In addition, abnormal reports tended to have more variability in the number of unique word phrases, as shown by the standard deviation that is nearly twice that of the normal reports.

Table 1. Number of unique phrases for each report

Normal Reports	**Abnormal Reports**
18	26
15	63
11	38
14	43
16	29
0	45
23	22
--	27
Avg: 13.857	**Avg:** 36.625
Std Dev: 7.151	**Std Dev:** 13.553

Further investigation into the word phrases of the abnormal reports revealed a wide-ranging vocabulary and semantics. Table 2 shows example word phrases from both normal and abnormal reports.

Table 2. Sample word phrases from reports

Normal Reports	Abnormal Reports
benign biopsy	intraductal carcinoma
breasts unchanged	rod shaped calcifications
microcalcifications identified	defined hyperdense nodule
remain unchanged	hypoechoic lesion
small deformable cyst	recommend excisional biopsy
benign macrocalcification	lobulated hypoechoic mass

Analysis of the word phrases provides further evidence to support our hypothesis that abnormal reports consists of statistically rare or unusual words, and thereby making them easier to identify in a large collection of reports.

After this first test, the MVS-GA was enhanced to extract s-grams from the individuals in the population. The s-grams discovered by the MVS-GA on the entire data set are shown in Tables 5 and 6. Table 5 shows the top ten most frequently occurring s-grams from the best solution obtained by the MVS-GA. These s-grams tend to uniquely define abnormal reports. Many of these s-grams refer to procedures that are performed in the event that a suspicious feature in the patient was observed by the radiologist. For example, the patient may be asked to return with a few weeks for additional imaging such as an ultrasound and magnification imaging. In addition, patients with suspicious features may undergo biopsy, and in some cases, may also have a needle localization performed to enhance the biopsy procedure. Furthermore, since breast cancer often affect the lymph nodes, radiologist look for abnormalities relating to the lymph nodes as well. As can be seen in Table 5, the MVS-GA successfully learned key s-grams that would significantly enhance automated classification of abnormal reports.

Table 5. Top ten most frequently occurring s-grams from best solution obtained by MVS-GA

Rank	S-gram	Example Phrase	Number of Variations Observed
1	magnification & views	magnification views requested	660
2	core & biopsy	stereotactic guided core biopsy of microcalcifications	633
3	needle & localization	ultrasound-guided needle localization procedure	245
4	nodular & density	showing questionable increased nodular density	2726
5	lymph & node	atypically located intramammary lymph node	748
6	needle & procedure	stereotactic needle core biopsy procedure	57
7	compression & views	right anterior compression views	772

8	spot & views	recommended utilizing spot views	852
9	spot & compression	spot compression image	1123
10	spot & magnification	medially exaggerated right cc spot magnification	650

Table 6 shows the top ten most frequently occurring s-grams that begin with the “no” and were learned from the failed individuals (i.e., individuals that were not selected for reproduction) in the MVS-GA. As discussed previously, most normal reports contain some form of a “negation” phrase. These phrases refer to the non-existence of a particular feature or condition in which the radiologist was searching. Abnormal reports may contain such negation phrases, however, abnormal reports tend to be more focused on the abnormalities that were found and not the abnormalities that were not found. Consequently, MVS-GA successfully learned from the failed samples the key s-grams of normal reports.

Table 6. Top ten most frequently occurring s-grams with the word "no"

Rank	S-gram	Example Phrase	Number of Variations Observed
1	no & suspicious	no finding strongly suspicious	1225
2	no & calcifications	no clear cut clustered punctate calcifications	137
3	no & evident	no mass lesions evident	46
4	no & masses	no new focal masses	365
5	no & malignancy	no specific evidence of malignancy	286
6	no & residual	no residual microcalcifications	56
7	no & skin	no skin abnormalities noted	68
8	no & thickening	no skin thickening seen	42
9	no & complications	no apparent complications	16
10	no & change	no apparent interval change	384

After the most frequent s-grams were extracted for normal and abnormal reports using the MVS-GA, these s-grams were then analyzed for their ability to distinguish between the two classes of reports. For this analysis the data set used consisted of 100 reports that were classified as “Normal” and 100 reports classified as “Abnormal” by a human expert. The average similarity (using the cosine similarity measure for all similarity calculations) between the reports within the Normal was computed. This same average was also computed for reports within the Abnormal class. Both of these averages are referred to as “Within-class Similarity”. Next, the average similarity of reports between the two classes was computed and referred to as “Between-class Similarity”. An ideal classifier should have a within-class similarity near 1.0 for each class and a between-class similarity near

0.0 for each pair of classes. Such a classifier would very accurately distinguish reports as being of a particular class.

Table 7 shows the results of using the most frequent normal and abnormal s-grams from Tables 5 and 6 to represent the report content. As can be seen in the table, the ability to distinguish the normal and abnormal reports is very low when using all of the terms in the reports. This is due to the amount of noise in the language of the reports. However, when using only the normal and abnormal s-grams that exist in the reports, the improvement is considerable, especially for the normal class. Again, this is due to the fact that normal reports are shorter and less variable in their language, while abnormal are longer and more ambiguous in their language. Also notable is that this improvement in the separability between normal and abnormal reports were achieved using considerably fewer terms, as can be seen by the average feature vector lengths. Approximately 90% of the terms could be removed from each report while providing significant improvements in the within-class and between-class similarities.

Table 7. Abnormal and Normal class separability using most frequent extracted s-grams

	Using all terms	Using only Abnormal and Normal S-Grams	Percent Change
Average "Normal" Feature Vector Length	57.79	5.81	-89.94%
Average "Abnormal" Feature Vector Length	117.79	9.57	-91.88%
Within-class similarity "Normal"	.1118 / .1114 (avg / std dev)	.8669 / .1070 (avg / std dev)	675.40%
Within-class similarity "Abnormal"	.1306 / .1056 (avg / std dev)	.3621 / .2999 (avg / std dev)	177.26%
Between-class similarity	.0618 / .0876 (avg / std dev)	.0531 / .1544 (avg / std dev)	-14.08%

8. SUMMARY

Currently, text analysis of mammography reports remains a significant challenge. However, solving this issue would provide numerous benefits. The work described here represents results in applying a GA to assist with identifying abnormal mammography reports from a large set of reports. Results were very encouraging and show tremendous potential for future work. Future work will seek to leverage this technique to develop a more advanced and specific training set of images to further enhance image-based algorithms.

9. ACKNOWLEDGMENTS

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

10. REFERENCES

[1] Abdalla, R. M., and Teufel, S., "A bootstrapping approach to unsupervised detection of cue phrase variants." In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (Sydney, Australia). COLING 2006. ACM Press, New York, NY, 2006.

[2] American College of Radiology (ACR). ACR BI-RADS® - Mammography. 4th Edition. In: ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas. Reston, VA. American College of Radiology; 2003.

[3] Cheng, W., Greaves, C. and Warren, M. 2006. "From n-gram to skipgram to concgram." International Journal of Corpus Linguistics 11/4: 411–33.

[4] X. Cui and T. E. Potok, "A distributed agent implementation of multiple species Flocking model for document partitioning clustering", in Lecture Notes in Computer Science. Vol. 4149 NAI Edinburgh, United Kingdom: Springer Verlag, Heidelberg, D-69121, Germany, 2006, pp. 124-137.

[5] Dreyer, K.J., Kalra, K.M., Maher, M.M., et al, "Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study", Radiology 324, 2 (Feb. 2005), 323-329.

[6] Dridi, O.; Ben Ahmed, M., "Building an Ontology-Based Framework For Semantic Information Retrieval: Application To Breast Cancer," Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on , pp.1-6, 7-11 April 2008.

[7] Duh, K., and Kirchhoff, K., "Automatic learning of language model structure". In Proceedings of the 20th International Conference on Computational Linguistics (Geneva, Switzerland). COLING 2004. ACM Press, New York, NY, 2004.

[8] Fickenscher, K.M., The New Frontier of Data Mining, Health Management Technology (26) 10:32-36, October 2005.

[9] C. Fox, "Lexical analysis and stoplists." In Information Retrieval: Data Structures and Algorithms (ed. W.B. Frakes and R. Baeza-Yates), Englewood Cliffs, NJ: Prentice Hall, 1992.

[10] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[11] Jing-Yan Wang; Zhen Zhu, "Framework of multi-agent information retrieval system based on ontology and its application," Machine Learning and Cybernetics, 2008 International Conference on , pp.1615-1620, 12-15 July 2008.

[12] Kai Kang; Kunhui Lin; Changle Zhou; Feng Guo, "Domain-Specific Information Retrieval Based on Improved Language Model," Fuzzy Systems and Knowledge

Discovery, 2007. FSKD 2007. Fourth International Conference on , pp.374-378, 24-27 Aug. 2007.

[13] H. Muehlenbein, “Parallel Genetic Algorithms, Population Genetics, and Combinatorial Optimization”, Proc. of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, 1989.

[14] P.P. Mutalik, et al., “Solving Combinatorial Optimization Problems Using Parallel Simulated Annealing and Parallel Genetic Algorithms”, Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing: technological challenges of the 1990's, 1992, pp 1031 – 1038.

[15] M. Nowostawski, and R. Poli, “Parallel Genetic Algorithm Taxonomy”, Proc. of the Third International Conference on Knowledge-based Intelligent Information Engineering Systems, 1999, pp. 88 – 92.

[16] M.Q. Patton, Qualitative Evaluation and Research Methods, Second Edition. Newbury Park, CA: Sage Publications, Inc., 1990

[17] R.M. Patton, and T.E. Potok, "Adaptive Sampling of Text Documents," Proc. of the 13th International Conference on Intelligent and Adaptive Systems and Software Engineering, July 2004.

[18] R. M. Patton and T. E. Potok, “Characterizing large text corpora using a maximum variation sampling genetic algorithm”, in Proc. of the 8th annual conference on Genetic and Evolutionary Computation Seattle, Washington, USA: ACM Press, 2006.

[19] Pirkola, A, Keskustalo, H., Leppänen, E., Känsälä, A.and Järvelin, K., "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants." Information Research, 7(2), 2002 [Available at http://InformationR.net/ir/7-2/paper126.html]

[20] M. Porter, “An algorithm for suffix stripping.” Program vol. 14, pp. 130-137, 1980.

[21] Porter Stemming Algorithm. Current Jan. 30, 2004. http://www.tartarus.org/~martin/PorterStemmer/

[22] V.V. Raghavan and S.K.M. Wong, “A critical analysis of vector space model for information retrieval.” Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, 1986.

[23] J.W. Reed, Y. Jiao, T.E. Potok, B.A. Klump, M.T. Elmore, and A.R. Hurson, “TF-ICF: A new term weighting scheme for clustering dynamic data streams”, In Proc. of the 5th International Conference on Machine Learning and Applications (ICMLA’06), pp.258-263, 2006.

[24] Reed, J.W., Potok, T. E., and Patton, R.M. 2004. "A multi-agent system for distributed cluster analysis," in Proceedings of Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS'04) Workshop in

conjunction with the 26th International Conference on Software Engineering Edinburgh, Scotland, UK: IEE, pp. 152-5.

[25] Rudolph, G., "Convergence analysis of canonical genetic algorithms," Neural Networks, IEEE Transactions on , vol.5, no.1, pp.96-101, Jan 1994.

[26] G. Salton, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[27] F. Sebastiani, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, Volume 34 , Issue 1 (March 2002) , pp 1 – 47.

[28] Siddiqui, T.J., "Integrating notion of agency and semantics in information retrieval: an intelligent multi-agent model," Intelligent Systems Design and Applications, 2005. ISDA '05. Proceedings. 5th International Conference on , pp. 160-165, 8-10 Sept. 2005.

[29] Starren, J., and Johnson, S.M., “Expressiveness of the Breast Imaging Reporting and Database System (BI-RADS)”, in Proc. of AMIAA Annual Fall Symp., 1997.

[30] R. Tanese, Distributed Genetic Algorithms for Function Optimization, Ph.D. thesis, University of Michigan, 1989, Computer Science and Engineering.

[31] Thomas, BJ; Ouellette, H; Halpern, EF; Rosenthal, DI. “Automated Computer-Assisted Categorization of Radiology Reports”, AJR: 184, 687-690. February 2005.

[32] S.K. Thompson, Sampling. John Wiley and Sons, Inc., New York, 1992.

[33] S.K. Thompson and G.A.F. Seber, Adaptive Sampling. John Wiley and Sons, Inc., New York, 1996.

[34] P. Yan, Y. Jiao, A.R. Hurson, and T.E. Potok, “Semantic-based information retrieval of biomedical data”, In Proc. of the 21st Annual ACM Symposium on Applied Computing – Semantic-Based Resource Discovery, Retrieval and Composition (RDRC), pp. 1700-1704, 2006.