



Final Report on SIG Analysis of Sibert Data

**Munitions Management Projects
Project MM-200501**

**Lawrence Carin, Levi Kennedy,
Xianyang Zhu, Yijun Yu and David Williams
Signal Innovations Group, Inc.**

October 15, 2008

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (15-07-2008)		2. REPORT TYPE Final Report		3. DATES COVERED (From March 2007 – March 2008 To)	
4. TITLE AND SUBTITLE Final Report on SIG Analysis of Sibert Data				5a. CONTRACT NUMBER W912HQ-05-C-0014-P00006	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Lawrence Carin, Levi Kennedy, Xianyang Zhu, Yijun Yu and David Williams				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Signal Innovations Group 1009 Slater Rd. Suite 200 Durham, NC 27703				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) ESTCP Environmental Security Technology Certification Program 901 North Stuart Street, Suite 303 Arlington, VA 22203				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT In this report we provide a summary of Signal Innovation Group's analysis of the UXO data collected at the Sibert test site. We discuss the feature extraction that has been performed, providing an examination of the features of UXO and non-UXO as a function of sensor type, and discuss how the final call lists were generated for submission to ESTCP. The data inversion process is also detailed. We provide a comprehensive report on the performance of the algorithms on the data, for passive and active learning, and across all sensors considered.					
15. SUBJECT TERMS UXO, cleanup					
16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT None	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)

CONTENTS

List of Figures	iv
List of Acronyms	xiii
Executive Summary	1
1 Introduction	2
1.1 Background	2
1.2 Objectives of the ESTCP UXO Discrimination Study	2
1.3 Technical objectives of the Discrimination Study	2
1.4 Regulatory Drivers and Stakeholder Issues	3
1.5 Management and Staffing	3
1.6 Specific Objective of Demonstration	4
1.7 Test site	4
2 Technology Description: Data Modeling and Inversion	6
2.1 Magnetometer model	6
2.2 EMI frequency-domain model	6
2.3 EMI time-domain model	8
3 Inversion of Sibert data	9
3.1 GEM3 data	9
3.2 EM63	10
3.3 EM61	14
4 Technology Description: Classifiers and Feature Selection	17
4.1 Supervised classifier	17
4.2 Semi-supervised classifier	18
4.3 Active learning	19
5 Cost, Performance and Technology Limitations	20
5.1 Factors Affecting Cost and Performance	20
5.2 Advantages and Limitations of the Technology	20

6	Algorithmic Details for Sibert Data	22
6.1	Feature Selection and Threshold Settings	22
6.2	Detailed Aspects of the Analysis	23
6.2.1	Analysis decisions	23
6.2.2	Parameters Estimated	24
6.2.3	Setting Thresholds	26
6.2.4	Analysis of GPO Data	27
6.3	Feature selection/weighting	28
7	Details on Setting Thresholds	30
7.1	Supervised vs semi-supervised learning and classification confidence .	30
7.2	Leave-one-out analysis of Sibert data	33
7.3	Setting of thresholds for blind test	41
8	Items Excluded from Classification Study	42
8.1	Subjective data removal	42
8.2	Details on removal methodology based on data inspection	45
9	Performance Assessment and Cost Assessment	50
9.1	Performance criteria	50
9.2	Performance confirmation methods	50
9.3	Data analysis, interpretation and evaluation	50
9.4	Cost reporting	51
9.5	Cost analysis	51
10	Classification Results - Non-Active Learning	52
10.1	Presentation format for all IDA-generated ROCs	52
10.2	EM61 sensor	53
10.3	Magnetometer	56
10.4	EM63 sensor	58
10.5	Concatenation of EM63 and magnetometer features	61
10.6	Concatenation of EM61 and magnetometer features	63

10.7	GEM3 sensor	65
10.8	Concatenation of GEM3 and magnetometer features	67
11	Active-Learning Classification Results	69
11.1	Intersection of EM61 and magnetometer data	69
11.2	Individual EM61 and magnetometer processing (post analysis)	71
12	Cost Assessment	73
12.1	Cost breakdown	73
12.2	Cost Benefit	74
	References	76

LIST OF FIGURES

1 Grid of measurements for GEM3 data. 9

2 Example measured GEM3 data. 10

3 Modeled results compared with measured data (the horizontal axis corresponds to different sensor positions). (a) $f=330$ Hz, real part; (b) $f=1470$ Hz, real part; (c) $f=330$ Hz, imaginary part; (d) $f=1470$ Hz, imaginary part 11

4 Example measured EM63 data (first time gate) and the boxed region of high-SNR data employed for inversion. 12

5 Example of the spatially sampled EM63 data points and those used in the inversion. 12

6 Example fits (time-gate one) for the EM63 data (a) measured, (b) model fit. 13

7 Example of a “bad” measured EM63 signature (time-gate one) 13

8 Example fit for a “bad” measured EM63 signature (time-gate one). (a) measured data, (b) model fit 14

9 Sample points for the EM61 in two modes. 15

10 Histogram of the EM61 features for targets and clutter within Sibert study. (a) dipole moment 1; (b) resonant frequency 1; (c) dipole moment 2; and (d) resonant frequency 2 15

11 Features extracted from the EM61 sensor, for *labeled* measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO. 23

12	Features extracted from the magnetometer sensor, for <i>labeled</i> measured at the Sibert site. The features are the dipole moment and model-fit error (from top to bottom, and left to right). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.	24
13	Features extracted from the EM63 sensor, for <i>labeled</i> measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.	25
14	Features extracted from the GEM3 sensor, for <i>labeled</i> measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.	26
15	Weights on the vector θ as computed for the EM61 sensor, using a supervised classifier.	29

- 16 The probability of being a UXO, $p(l = 1|\mathbf{x})$ is plotted in a two-dimensional feature space characteristic of two of the features in the overall feature vector \mathbf{x} . Results are shown for the EM63 sensor, using the labeled and unlabeled data from the Sibert site. In this plot a supervised classifier is considered, and therefore $p(l = 1|\mathbf{x})$ is designed using only the labeled data. An important thing to note is how confident the classifier is in the decision boundary: above the boundary for which $p(l = 1|\mathbf{x}) = 0.5$ one observes that $p(l = 1|\mathbf{x}) \approx 1$ very quickly, and below $p(l = 1|\mathbf{x}) = 0.5$ we observe $p(l = 1|\mathbf{x},) \approx 0$ very quickly as a function of \mathbf{x} . Hence, based on the limited available labeled data, and in absence of the context provided by the unlabeled data, the classifier is very confident in what parts of feature space \mathbf{x} correspond to UXO and non-UXO. The classifier is designed based on all EM63 features, and the plot here considers the classifier decision as viewed in a two-dimensional plane within that feature space (corresponding to the decay constant and moment associated with the first EMI dipole). 31
- 17 The probability of being a UXO, $p(l = 1|\mathbf{x})$ is plotted in a two-dimensional feature space characteristic of two of the features in the overall feature vector \mathbf{x} . Results are shown for the EM63 sensor, using the labeled and unlabeled data from the Sibert site. In this plot a semi-supervised classifier is considered, and therefore $p(l = 1|\mathbf{x})$ is designed using both the labeled *and* unlabeled data. In comparison to Figure 16 note that the $p(l = 1|\mathbf{x}) = 0.5$ boundary is shifted slightly; of more importance, note the far more gradual change in the probabilities $p(l = 1|\mathbf{x})$ for features \mathbf{x} away from the region $p(l = 1|\mathbf{x}) = 0.5$. This implies that the semi-supervised classifier is less confident in which features \mathbf{x} correspond to UXO, as a result of the context provided by the unlabeled data. The classifier is designed based on all EM63 features, and the plot here considers the classifier decision as viewed in a two-dimensional plane within that feature space (corresponding to the decay constant and moment associated with the first EMI dipole). 32
- 18 Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM61 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted. 33

19	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM61 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	35
20	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	35
21	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted. . . .	36
22	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM63 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	36
23	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM63 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	37
24	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled GEM3 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	37
25	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled GEM3 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	38
26	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and EM61 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	38
27	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and EM61 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	39

28	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and EM63 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	39
29	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and EM63 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	40
30	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and GEM3 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.	40
31	Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and GEM3 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.	41
32	A summary of the number of signatures per sensor, the number of labeled training examples given per sensor, the total number of excluded items, and the number of these that came from the training set. None of the items excluded from the EM61 and magnetometer sensors was UXO.	43
33	Representative example data that were excluded from the study from the magnetometer sensor.	44
34	Representative example data that were excluded from the study from the EM61 sensor.	44
35	Example EM61 signatures that were deemed of high-enough quality to perform classification	45
36	Example measured EM61 signatures (top) and associated model reconstruction (bottom). The left-most example was removed from the classification analysis, and the middle and right data were used within the classifier. The white point in each figure is the as-given target location.	46

37	Example measured EM61 signatures that were removed because of overlapping signatures. Note that the bottom two correspond to very-large signatures near the given target location (white dot), and we could not be sure if this was a location error or a strong nearby signature. This class corresponded to 30% of the removed cases.	47
38	Example measured EM61 signatures that were removed because it was confusing as to what precisely was the signature (highly anomalous signatures). This class corresponded to 55% of the removed cases.	47
39	Example measured EM61 signatures that were removed because of weak signals in the location of the specified target location (white point). Note that there are sometimes strong nearby targets and it is not clear if there is actually target-location error. This class corresponded to 15% of the removed cases.	48
40	Example measured magnetometer signatures that were removed because of overlapping signatures. Note that the bottom two correspond to very-large signatures near the given target location (white dot), and we could not be sure if this was a location error or a strong nearby signature. This class corresponded to 39% of the removed cases.	48
41	Example measured magnetometer signatures that were removed because it was confusing as to what precisely was the signature (highly anomalous signatures). This class corresponded to 7% of the removed cases.	49
42	Example measured magnetometer signatures that were removed because of weak signals in the location of the specified target location (white point). Note that there are sometimes strong nearby targets and it is not clear if there is actually target-location error. This class corresponded to 54% of the removed cases.	49
43	Performance criteria for Sibert test.	50
44	Receiver operating characteristics (ROCs) for the Sibert site, based on the EM61 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	54
45	Receiver operating characteristics (ROCs) for the Sibert site, based on the EM61 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	55

46	Summary Sibert performance for the EM61 sensor, using a semi-supervised classifier.	56
47	Summary Sibert performance for the EM61 sensor, using a supervised classifier.	56
48	Receiver operating characteristics (ROCs) for the Sibert site, based on the magnetometer sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	57
49	Receiver operating characteristics (ROCs) for the Sibert site, based on the magnetometer sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	57
50	Summary Sibert performance for the magnetometer sensor, using a semi-supervised classifier.	58
51	Summary Sibert performance for the magnetometer sensor, using a supervised classifier.	58
52	Receiver operating characteristics (ROCs) for the Sibert site, based on the EM63 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	59
53	Receiver operating characteristics (ROCs) for the Sibert site, based on the EM63 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	60
54	Summary Sibert performance for the EM63 sensor, using a semi-supervised classifier.	60
55	Summary Sibert performance for the EM63 sensor, using a supervised classifier.	61
56	Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM63 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	61
57	Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM63 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	62
58	Summary Sibert performance for the concatenation EM63 and magnetometer features, using a semi-supervised classifier.	62
59	Summary Sibert performance for the concatenation EM63 and magnetometer features, using a supervised classifier.	63

60	Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM61 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	64
61	Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM61 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	64
62	Summary Sibert performance for the concatenation EM61 and magnetometer features, using a semi-supervised classifier.	65
63	Summary Sibert performance for the concatenation EM61 and magnetometer features, using a supervised classifier.	65
64	Receiver operating characteristics (ROCs) for the Sibert site, based on the GEM3 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	66
65	Receiver operating characteristics (ROCs) for the Sibert site, based on the GEM3 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	66
66	Summary Sibert performance for the GEM3 sensor.	67
67	Receiver operating characteristics (ROCs) for the Sibert site, based on concatenation of the GEM3 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.	68
68	Receiver operating characteristics (ROCs) for the Sibert site, based on concatenation of the GEM3 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.	68
69	Receiver operating characteristics (ROCs) for the Sibert site, based on concatenated features from the EM61 and magnetometer sensors. The results are for a semi-supervised classifier, and are based on labeled data acquired via active learning. . .	70
70	Receiver operating characteristics (ROCs) for the Sibert site, based on concatenated features from the EM61 and magnetometer sensors. The results are for a supervised classifier, and are based on labeled data acquired via active learning.	70

71	Summary Sibert performance for concatenated features from the EM61 and magnetometer sensors, based on a semi-supervised classifier. The labeled data were acquired via active learning.	71
72	Summary Sibert performance for concatenated features from the EM61 and magnetometer sensors, based on a supervised classifier. The labeled data were acquired via active learning.	71
73	Receiver operating characteristics (ROCs) for the EM61 (left) and magnetometer (right) sensors alone. The results are for a supervised classifier (the ROCs for a semi-supervised classifier are virtually identical), and are based on labeled data acquired via active learning.	72
74	Cost model for work performed during the Camp Sibert discrimination study. Active learning was only performed on a subset of the sensor data provided.	75

List of Acronyms

CPU: Central processor unit

EMI: Electromagnetic induction

FUDS: Formerly used defense site

GOF: Goodness of fit

GPO: Geophysical prove-out

IDA: Institute for Defense Analyses

MAG: Magnetometer

NRL: Naval Research Laboratory

ROC: Receiver operating characteristic

SIG: Signal Innovations Group

UXO: Unexploded ordnance

Executive Summary

In this report we provide a summary of Signal Innovation Group's analysis of the UXO data collected at the Sibert test site. We discuss the feature extraction that has been performed, providing an examination of the features of UXO and non-UXO as a function of sensor type, and discuss how the final call lists were generated for submission to ESTCP. The data inversion process is also detailed. We provide a comprehensive report on the performance of the algorithms on the data, for passive and active learning, and across all sensors considered.

I. Introduction

A. Background

In 2003, the Defense Science Board observed: “The problem is that instruments that can detect the buried UXOs also detect numerous scrap metal objects and other artifacts, which leads to an enormous amount of expensive digging. Typically 100 holes may be dug before a real UXO is unearthed! The Task Force assessment is that much of this wasteful digging can be eliminated by the use of more advanced technology instruments that exploit modern digital processing and advanced multi-mode sensors to achieve an improved level of discrimination of scrap from UXOs.” Significant progress has been made in discrimination technology. To date, testing of these approaches has been primarily limited to test sites with only limited application at live sites. Acceptance of discrimination technologies requires demonstration of system capabilities at real UXO sites under real-world conditions. Any attempt to declare detected anomalies to be harmless and requiring no further investigation will require demonstration to regulators of not only individual technologies, but of an entire decision making process.

B. Objectives of the ESTCP UXO Discrimination Study

As outlined in the Environmental Security Technology Certification Program (ESTCP) Unexploded Ordnance (UXO) Discrimination Study Demonstration Plan, the objectives of the study were twofold. First, the study was designed to test and validate UXO detection and discrimination capabilities of currently available and emerging technologies on real sites under operational conditions. Second, the ESTCP Program Office and their demonstrators have investigated, in cooperation with regulators and program managers, how UXO discrimination technologies may be implemented in cleanup operations.

C. Technical objectives of the Discrimination Study

The study was designed to test and evaluate the capabilities of various UXO discrimination processes, each consisting of a selected sensor hardware system, a survey mode, and a software-

based processing step. These advanced methods are compared to existing practices with the goal of validating the pilot technologies for the following:

- Detection of UXOs
- Identification of features that can help distinguish scrap and other clutter from UXO
- Reduction of false alarms (items that could be safely left in the ground that are incorrectly classified as UXO) while maintaining acceptable Pd's
- Quantification of the cost and time impact of advanced methods on the overall cleanup process as compared to existing practices

Additionally, the study aimed to understand the applicability and limitations of the selected technologies in the context of project objectives, site characteristics, and suspected ordnance contamination. Sources of uncertainty in the discrimination process have been identified and their impact quantified to support decision making. This includes issues such as the impact of data quality due to how the data are collected. The process for making the dig/no-dig decision process is explored. Potential QA/QC processes for discrimination are also explored. Finally, high-quality, well documented data was collected to support the next generation of signal processing research.

D. Regulatory Drivers and Stakeholder Issues

ESTCP assembled an Advisory Group to address the regulatory, programmatic, and stakeholder acceptance issues associated with the implementation of discrimination in the Munitions Response (MR) process.

E. Management and Staffing

The demonstration summarized here was conducted with the support of several SIG personnel. Dr. Lawrence Carin (PI) acted as the Quality Assurance (QA) Officer, and also managed the demonstration process and reporting. Dr. Xianyang Zhu, Mr. Levi Kennedy, Dr. Yijun Yu, and Dr. David Williams performed the data processing and analysis. Dr. Paul Runkle provided cost management and general oversight.

F. Specific Objective of Demonstration

The purpose of this demonstration was to apply and evaluate the classification algorithms on the Camp Sibert data set to demonstrate that some non-UXO items can be classified correctly and hence left in the ground, while maintaining a given level of detection performance. The performance of seven different sensor combinations were compared. Moreover, for each combination, after the labeled data are defined for training, classification results were obtained using two different classification approaches. The first approach employed a supervised classifier, using only the labeled training data, not accounting for the context provided by the unlabeled data. The second approach is semi-supervised, exploiting the unlabeled data in addition to the labeled data when building the classifier (details of these approaches are provided below). In addition to using the labeled data as provided by ESTCP to design the above algorithms, for one set of sensor combinations we employed active learning to define the set of signatures for which acquisition of the associated labels were most informative for classifier design (we only considered one sensor combination for active learning because it was likely that the different sensors may imply different items to be informative if labeled, and hence by considering many different sensors when performing active learning, it was feared that too much of the site will be excavated for acquisition of labeled data). In the proposed analysis features were extracted from the sensor data, employing magnetometer and induction models developed at Duke. For each approach described above, a dig list was constructed to order the anomalies based on the probability of being UXO. ROC curves were then constructed for each method based on these lists.

G. Test site

The ESTCP UXO Discrimination Study Demonstration site has been selected to be Camp Sibert, Alabama. The land, which is under private ownership and is used as a hunting camp, is located within the boundaries of Site 18 of the former Camp Sibert FUDS. Information on the Camp Sibert FUDS is available in the archival literature such as an Archives Search Report (ASR) developed in 1993. The former Camp Sibert is located in the Canoe Creek Valley between Chandler Mountain and Red Mountain to the northwest, and Dunaway Mountain and Canoe Creek Mountain to the southeast. Camp Sibert is comprised of mainly sparsely inhabited farmland and woodland and encompasses approximately 37,035 acres. The City of Gadsden is growing

towards the former camp boundaries from the north. The Gadsden Municipal Airport occupies the former Army airfield in the northern portion of the site. The site is located approximately 50 miles northwest of the Birmingham Regional Airport or 86 miles southeast of the Huntsville International Airport. The site is near exit 181 off of Interstate 59 in Gadsden and located approximately 8 miles southwest of the City of Gadsden, near the Gadsden Municipal Airport. The area that would become Camp Sibert was selected in the spring of 1942 for use in the development of a Replacement Training Center (RTC) for the Army Chemical Warfare Service. The RTC was moved from Edgewood, Maryland to Alabama in 1942. In the fall of 1942, the Unit Training Center (UTC) was added as a second command. Units and individual replacements were trained in aspects of both basic military training and in the use of chemical weapons, decontamination procedures, and smoke operations from 1942 to 1945. Mustard, phosgene, and possibly other agents were used in the training. This facility provided a previously unavailable opportunity for large scale training with chemical agent. Conventional weapons training was also conducted with several types and calibers fired, with the 4.2-inch mortar being the heavy weapon used most in training. The US Army also constructed an airfield for the simulation of chemical air attacks against troops. The camp was closed at the end of the war in 1945, and the chemical school transferred to Ft. McClellan, Alabama. The U.S. Army Technical Escort Unit (TEU) undertook several cleanup operations during 1947 and 1948; however, conventional ordnance may still exist in several locations. After decontamination of various ranges and toxic areas in 1948, the land was declared excess and transferred to private and local government ownership. A number of investigations have been conducted on various areas of the former Camp Sibert from 1990 to present. These investigations included record searches, interviews, surface assessments, geophysical surveys, and intrusive activities. The ESTCP UXO Discrimination Study Demonstration Site is located within the confines of Site 18, Japanese Pillbox Area No. 2, of the former Camp Sibert FUDS. Simulated pillbox fortifications were attacked first with WP ammunition in the 4.2-inch chemical mortars followed by troop advance and another volley of HE-filled 4.2-inch mortars. Assault troops would then attack the pillboxes using machine guns, flamethrowers, and grenades. The locations of nine possible bunkers and one trench in 1943 were identified as part of the 1999 TEC investigation. There is historical evidence of intact 4.2-inch mortars and 4.2-inch mortar debris being at the site. As part of the recent investigations, a geophysical survey of Site 18 has been conducted and multiple anomalies were identified.

II. Technology Description: Data Modeling and Inversion

Below we provide details of the models used for fitting the measured data, with the associated parameters used in the subsequent classifiers. A discussion is provided for each of the data types considered by SIG in the Sibert study.

A. Magnetometer model

For sensors sufficiently distant from the target (relative to the target dimensions, *i.e.*, the observation point is in the far field of the target), the magnetic vector potential may be represented approximately as

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_o}{4\pi} \nabla \frac{1}{R} \times \mathbf{m} \quad (1)$$

where \mathbf{m} is the magnetic dipole moment (the other components due to multipole expansion decay rapidly with distance and hence are neglected here), R is the distance between the target center and the observation point, and μ_o is the permeability of free space. Then the associated magnetic field can be expressed as

$$\mathbf{H} = \frac{1}{2\pi R^3} [3(\mathbf{m} \cdot \hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}] \quad (2)$$

where $\hat{\mathbf{r}}$ is a unit vector from the target center to the observation point. The magnetometer field measured as a function of position on a surface can then be fitted by the above dipole moment model. The parameters one may extract with this model are the target depth, the magnetic dipole orientation, and the dipole moment.

B. EMI frequency-domain model

The EMI response of simple targets can be represented in terms of a frequency dependent magnetic dipole, constituting a generalization of the magnetometer model. In particular, the magnetic dipole moment \mathbf{m} of a target can be represented as $\mathbf{m} = \mathbf{M} \cdot \mathbf{H}_{inc}$, where \mathbf{H}_{inc}

represents the incident magnetic field and \mathbf{M} is a tensor that relates the incident field and the dipole moment. For a UXO with axis along the z direction, the magnetization tensor may be expressed as:

$$\mathbf{M}(\omega) = \hat{z}\hat{z}[m_z(0) + \sum_k \frac{\omega m_{zk}}{\omega - j\omega_{zk}}] + (\hat{x}\hat{x} + \hat{y}\hat{y})[m_p(0) + \sum_i \frac{\omega m_{pi}}{\omega - j\omega_{pi}}] \quad (3)$$

where \hat{x} , \hat{y} and \hat{z} are unit vectors in the x , y , and z directions, respectively. The terms $m_z(0)$ and $m_p(0)$ account for the induced magnetization produced for ferrous targets, with these constants equal to zero for nonpermeable targets, and the terms in the summations account for the frequency dependent character. For simple targets, typically we only require the first term in each sum, representative of the principal dipole mode along each of the principal axes. Once the excitation magnetic field \mathbf{H}_{inc} is given, the dipole moment of the target can then be easily obtained according to the above magnetization tensor. Then the associated magnetic vector potential and the magnetic field at the observation point can be calculated readily as in the magnetometer model. If we assume that the EMI source responsible of \mathbf{H}_{inc} can be represented as a magnetic dipole with moment \mathbf{m}_s , then the incident magnetic field may be expressed as

$$\mathbf{H}_{inc} = \hat{\mathbf{r}}_{st} \frac{\mu_0}{2\pi} \frac{\mathbf{m}_s \cdot \hat{\mathbf{r}}_{st}}{R^3} \quad (4)$$

where $\hat{\mathbf{r}}_{st}$ is a unit vector directed from the source to the target center. If we assume that the source and observer coils are co-located, then the total magnetic field observed at the sensor can be represented by

$$\mathbf{H}_{rec} \approx \frac{\hat{\mathbf{r}}_{st}}{R^6} \hat{\mathbf{r}}_{st} \cdot \mathbf{U}^T \mathbf{M} \mathbf{U} \cdot \hat{\mathbf{r}}_{st} \quad (5)$$

where the proportionality constant depends on the strength of the dipole source \mathbf{m}_s and the characteristics of the receiver. The 3×3 unitary matrix \mathbf{U} rotates the fields from the coordinate system of the sensor to the coordinate system of the target, and \mathbf{U}^T transforms the dipole fields of the target back to the coordinate system of the observer (sensor).

Similarly, the parameters of the target depth, the target orientation, the dipole moments, and dipole frequencies (corresponding to decay constants) can be extracted from the measured data

based on this model.

Note that, in the above discussion, it was assumed that the target under test is rotationally symmetric, and therefore two of the dipole parameters were assumed to be the same (the x and y components, perpendicular to the z -directed rotation axis). It is recognized that many non-UXO may not satisfy this assumption of rotational symmetry (although it is anticipated that the confusing clutter are likely to). We may readily remove the assumption of rotation symmetry by treating all three components of the dipole moment as distinct. In previous studies at Duke the assumption of a rotationally-symmetric magnetization tensor has yielded good results, and therefore this assumption is used here. In a post-analysis step, SIG will revisit this assumption and examine its impact on performance.

C. EMI time-domain model

The EMI magnetic-dipole model in the time domain can be readily found via a Fourier transform of the above frequency domain model. Thus the pulse response can be expressed as

$$\begin{aligned} \mathbf{M}(t) = & \hat{z}\hat{z}[m_z(0)\delta(t) + \frac{\partial}{\partial t} \sum_k u(t)m_{zk}exp(-\omega_{zk}t)] \\ & + (\hat{x}\hat{x} + \hat{y}\hat{y})[m_p(0)\delta(t) + \frac{\partial}{\partial t} \sum_i u(t)m_{pi}exp(-\omega_{pi}t)] \end{aligned} \quad (6)$$

where $u(t)$ is a step function. The time-domain measured data collected by the EM61 and EM63 are fitted by the above model, and the corresponding parameters like the dipole moments, resonant frequencies, target depth and orientation may be extracted accordingly.

III. Inversion of Sibert data

A. GEM3 data

The GEM3 data considered by SIG were cued. That is, all the data were collected at fixed points above the target. For example, for some targets the map of the points is shown in Figure 1.

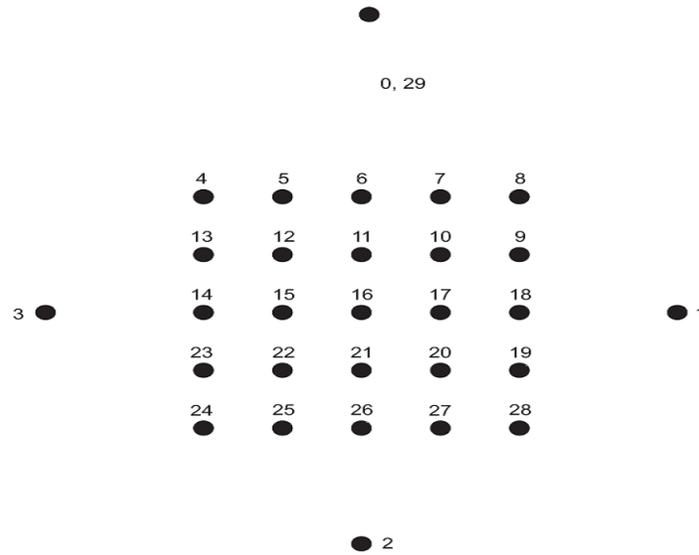


Fig. 1. Grid of measurements for GEM3 data.

We can see that there are 29 spatial sensor points. At each point except the top one (where the data were collected twice, at the beginning and end of the measurement) measured data were collected at 10 different frequencies for a short time. Only measured data collected at points 4-28 were used for the purpose of inversion, since the other points are used for calibration and were generally far from the targets. The 10 frequencies considered were 30, 90, 150, 330, 690, 1470, 3090, 6510, 13950, and 30030 Hz. At each frequency, multiple measurements were performed. Typical measured data are shown in Figure 2. We note that the measured data at frequency 30Hz is not stable; therefore, the measured data from this frequency were not be used to extract the model parameters.

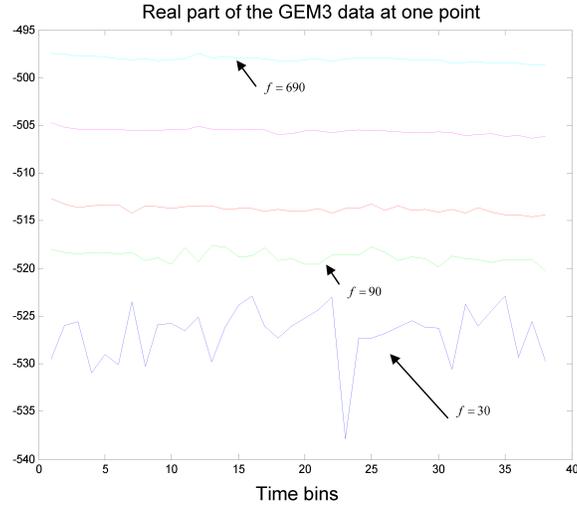


Fig. 2. Example measured GEM3 data.

For all the other remaining frequencies, we calculated the mean of the measured data at each point. Those data were then used for the purpose of inversion, which led to a system of nonlinear equations. The over-determined problem was solved by a least square method [1].

It should be noted that there typically exist many local-optimal solutions [1]; this is true for all the sensors and models considered. To find the global minimum (best-fit solution), we solved the non-linear least squares problems 64 times, with different random inversion initializations. The solution with the minimum fitting error was chosen as the final result. Typical GEM3 modeled results are shown in Figure 3, compared with measured data at frequency of 330 and 1470 Hz.

B. EM63

The EM63 sensor operates over a wide dynamic range of time samples (much wider than the EM61), providing a complete description of the time-decay (the transient response) associated with the target. The data were collected on 26 time gates, geometrically spaced in time from 180 microseconds to 25 milliseconds. On the other hand, much more CPU time is required for the data inversion since there is a significant quantity of measured data (in space and time). In order to reduce the CPU time, we preprocessed the measured data in three respects. First,

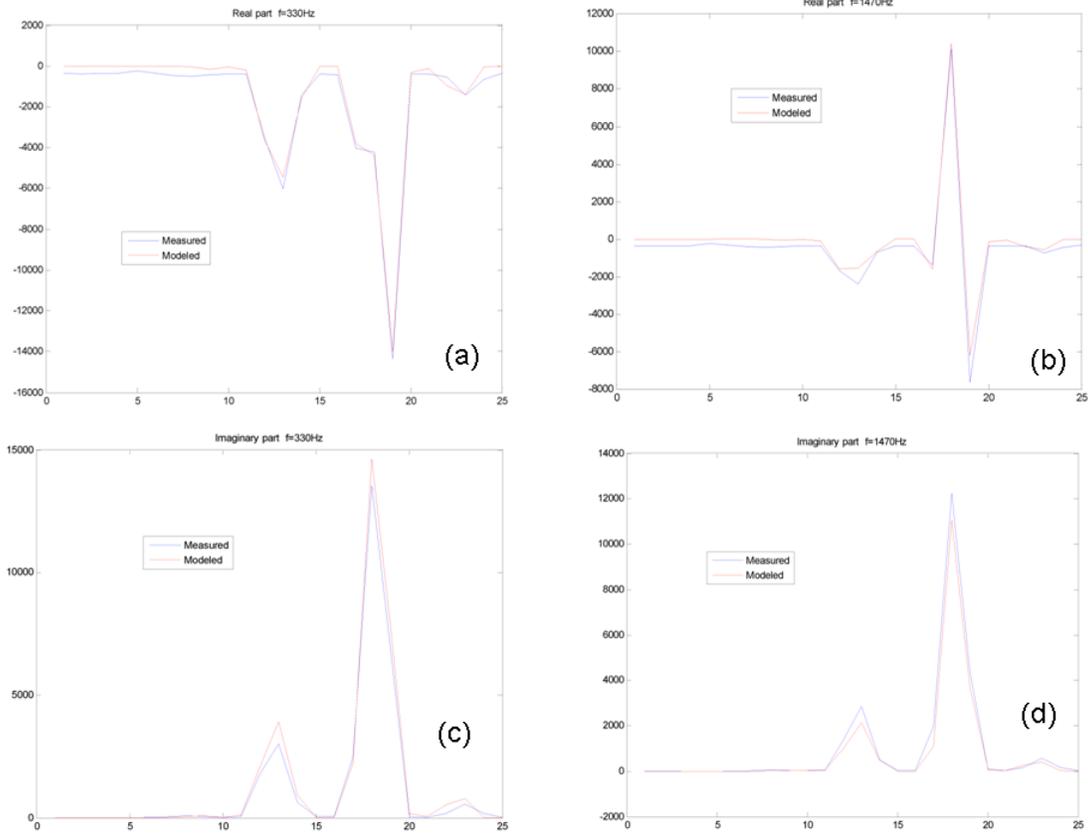


Fig. 3. Modeled results compared with measured data (the horizontal axis corresponds to different sensor positions). (a) $f=330$ Hz, real part; (b) $f=1470\text{Hz}$, real part; (c) $f=330$ Hz, imaginary part; (d) $f=1470$ Hz, imaginary part

we excluded measured data with low signal-noise ratio. This preprocessing was implemented manually by checking the image of the measured data. The high-SNR data defined a spatial box about the target, with which inversion was performed (see Figure 4).

Secondly, not all of the time gates were used. The measured data after a particular time gate (for example, time gate 15) are very small compared with the values at the first time gate and are deemed to be too noisy to be employed within the inversion. We studied the inversion results based on the measured data using a variable number of time gates, and found that using the measured data associated with the first 10 time gates gave better results from the view point of fitting error. This rule was applied for all the measured data collected by the EM63 sensor. If there are still too many measured data after removing later time gates, the original measured data

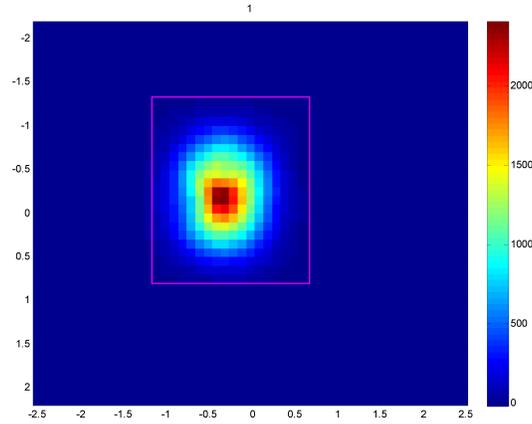


Fig. 4. Example measured EM63 data (first time gate) and the boxed region of high-SNR data employed for inversion.

were spatially subsampled so that the total number of points will be less than some specified value (this value is set to 200 in this effort). An example is shown in Figure 5, where the red dots represent the sampled data, which are a subset of the original measured data.

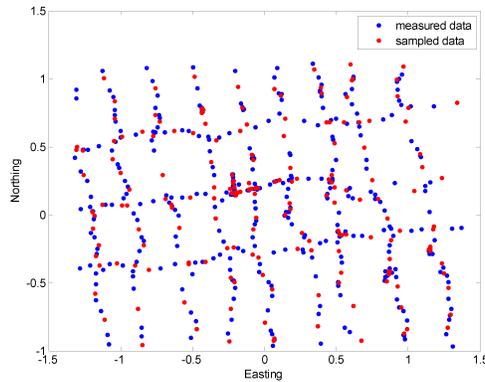


Fig. 5. Example of the spatially sampled EM63 data points and those used in the inversion.

Similarly to the data inversion for the sensor GEM3, 128 randomly-initialized least-square inversions were performed for each target, and the one with the minimum fitting error was chosen as the final solution. The comparison between the measured and modeled data images is shown in Figure 6. This method of addressing local-optimal least-squares inversions was also

employed for the magnetometer and EM61 data (discussed below).

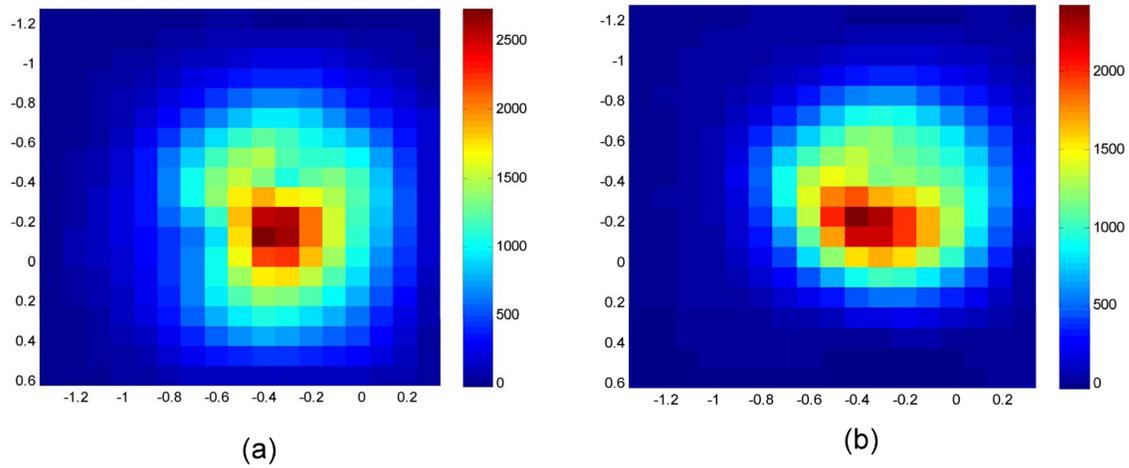


Fig. 6. Example fits (time-gate one) for the EM63 data (a) measured, (b) model fit.

To give an example of attempted inversion with “bad” measured data, a typical example (3rd GPO target) is shown in Figure 7. This poor target signature (for model inversion) may result from other surrounding targets.

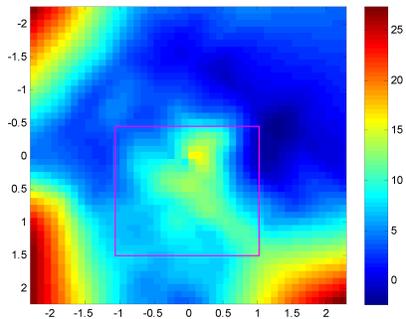


Fig. 7. Example of a “bad” measured EM63 signature (time-gate one)

For such “bad” data the model fits are poor (see Figure 8), and such data were not considered within the classification phase.

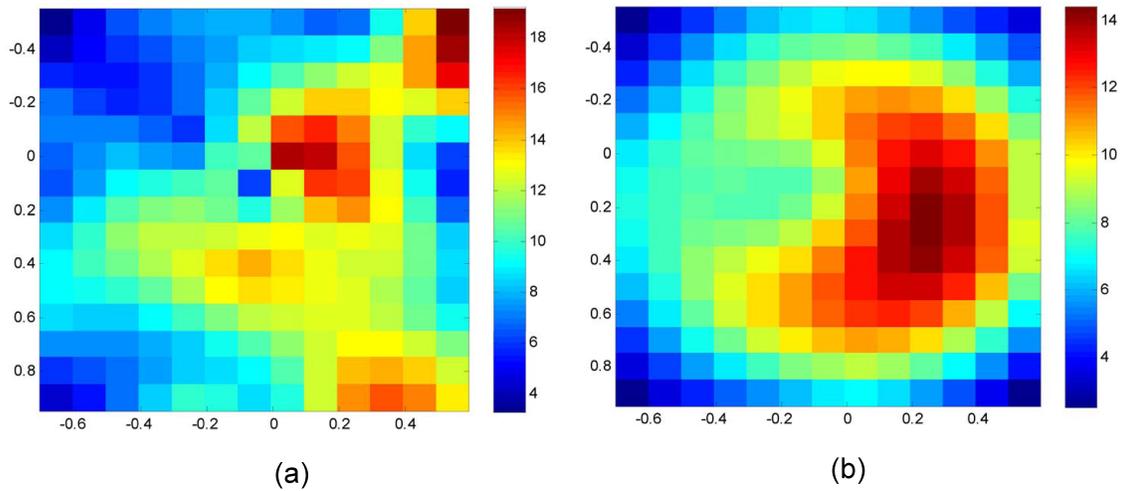


Fig. 8. Example fit for a “bad” measured EM63 signature (time-gate one). (a) measured data, (b) model fit

C. EM61

The process of data selection for inversion, pre-inversion data processing and the inversion process for the sensors MAG and EM61 are the same as those for EM63 (apart from the special time-bin issues associated with the long-decay EM63 data). The EM61 model fitting was based on four time gates; *after the test* we recognized that the EM61 was in “differential mode”, while we modeled it as being non-differential mode (see Figure 9). During the test we recognized somewhat anomalous behavior of gate two (not recognizing it was in differential mode), and we attempted inversion using gates 1, 3 and 4 as well as gates 1-4; the results were similar. In the post-mission analysis to follow we will refit the EM61 data, using the differential-mode parameters (treating gate 2 properly), but it is anticipated that performance changes will be minor – as indicated below, classification performance with the EM61 sensor was already relatively good (for the items on which classification was attempted).

The statistics of the inverted parameters based on the sensor EM61 are shown in Figure 10, where the histograms of the four parameters (two dipole moments plus two resonant frequencies) are presented for both UXOs and clutters; these histograms are for the entire Sibert site, based on truth (labels) provided after the test.

NRL EM61 MkII Array Gate Timing Parameters			
4 Gate Mode (Bottom Coil)	Delay (μ s)	Differential Mode	Delay (μ s)
Gate 1	307	Bottom Gate 1	307
Gate 2	508	Top Gate 1	307
Gate 3	738	Bottom Gate 2	738
Gate 4	1000	Bottom Gate 3	1000

Fig. 9. Sample points for the EM61 in two modes.

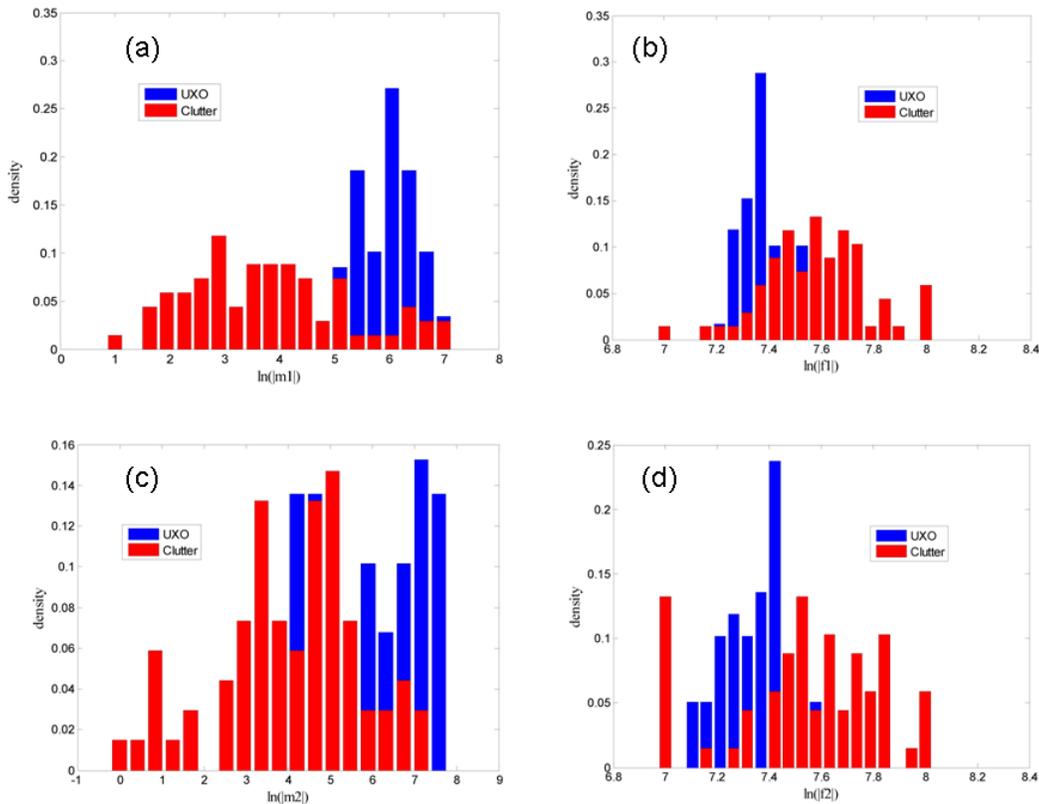


Fig. 10. Histogram of the EM61 features for targets and clutter within Sibert study. (a) dipole moment 1; (b) resonant frequency 1; (c) dipole moment 2; and (d) resonant frequency 2

The possible causes for spreads in the parameters are: (1) For some shallow-buried targets, the sensors are too close (the sensors are in the near field of the targets); therefore, the validity of the approximation will be broken. (2) As mentioned before, there are often local-optimal inversion solutions; this can be improved by increasing the number of random initializations. Concerning

the locations of the items, the variances of the inversion locations relative to ground truth in Easting, Northing and depth are 0.034, 0.037, and 0.110 meters.

IV. Technology Description: Classifiers and Feature Selection

We here provide a brief discussion of the supervised and semi-supervised classifiers; active learning is also briefly reviewed. This discussion is *not* meant to be complete, but rather is meant to provide a quick feel for the form of the algorithms employed in this study. The PI has written extensive papers on all of the methods utilized here, with references to that work cited below.

A. Supervised classifier

Assume that \mathbf{x} represents a feature vector we wish to classify. Further, assume that $\{\mathbf{x}_n, y_n\}_{n=1, N}$ represent a set of labeled (training) data, with which we wish to learn a classifier; the labels y_n are binary (1 or 0), corresponding to UXO/non-UXO. A sparse kernel classifier is constructed by considering a function

$$g(\mathbf{x}; \mathbf{w}) = w_0 + \sum_{n=1}^N w_n K(\mathbf{x}, \mathbf{x}_n) \quad (7)$$

where $K(\mathbf{x}, \mathbf{x}_n)$ is a “kernel” function, an example of which is the radial basis function $K(\mathbf{x}, \mathbf{x}_n) = \exp(\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}_n\|^2)$. The $N+1$ dimensional vector \mathbf{w} represents the set of weights $\{w_0, w_1, \dots, w_N\}$, and it is this vector we wish to learn based on the labeled data $\{\mathbf{x}_n, y_n\}_{n=1, N}$. The probability that the label $y = 1$ for feature vector \mathbf{x} is expressed in terms of a “logistic” link function

$$p(y = 1 | \mathbf{x}, \mathbf{w}) = \sigma[g(\mathbf{x}; \mathbf{w})] \quad (8)$$

where $\sigma(g) = \exp(g)/[1 + \exp(g)]$. Note that the link function maps a real number $-\infty < g(\mathbf{x}; \mathbf{w}) < \infty$ to a real number between zero and one (the latter constituting a probability distribution).

In the employed Bayesian analysis we define a prior $p(\mathbf{w} | \Gamma)$, parametrized by the hyperparameters Γ , that imposes the belief that most of the w_n should be zero, implying that only a subset of the training signatures are employed within the final classifier. Given the labeled data $\{\mathbf{x}_n, y_n\}_{n=1, N}$ and the prior $p(\mathbf{w})$ we may infer a posterior distribution on the weights \mathbf{w} :

$$p(\mathbf{w}|\{\mathbf{x}_n, y_n\}_{n=1,N}, \Gamma) \propto p(\mathbf{w}) \prod_{n=1}^N \{\sigma[g(\mathbf{x}_n; \mathbf{w})]\}^{y_n} \{\sigma[g(\mathbf{x}_n; \mathbf{w})]\}^{1-y_n} \quad (9)$$

Based on the inferred model-parameter posterior $p(\mathbf{w}|\{\mathbf{x}_n, y_n\}_{n=1,N}, \Gamma)$, which is designed based entirely on the labeled training data (thus the term “supervised” classifier), one may perform inference about the labels y for new (unlabeled) feature vectors \mathbf{x} observed during the testing phase [2].

B. Semi-supervised classifier

A key aspect of the above supervised classifier is that it does not exploit the available *contextual information* provided by the abundant unlabeled data at the site under test. In this sense, the supervised classifier does not utilize all available information. A semi-supervised classifier employs the available labeled data $\{\mathbf{x}_n, y_n\}_{n=1,N}$, while also utilizing the additional unlabeled data $\{\mathbf{x}_n\}_{n=N+1,N+M}$; we assume M unlabeled signatures from the site under test, and we wish to infer labels (the identity, UXO/non-UXO) for these signatures. Toward this end, we define a matrix \mathbf{K} , the (i, j) th element of which is $K(\mathbf{x}_i, \mathbf{x}_j)$, where $1 \leq i \leq N+M$ and $1 \leq j \leq N+M$, and the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ need not be the same one as used in the above classifier. The matrix \mathbf{K} defines similarities between all data points: if $K(\mathbf{x}_i, \mathbf{x}_j)$ is large then this implies that \mathbf{x}_i and \mathbf{x}_j are similar, while when $K(\mathbf{x}_i, \mathbf{x}_j)$ tends to zero the \mathbf{x}_i and \mathbf{x}_j are dissimilar.

We now normalize the rows of \mathbf{K} such that they sum to one. Specifically, the new normalized matrix \mathbf{A} has elements defined as

$$\mathbf{A}(i, j) = \frac{\mathbf{K}(i, j)}{\sum_{k=1}^{N+M} \mathbf{K}(i, k)} \quad (10)$$

The $(N+M) \times (N+M)$ matrix \mathbf{A} defines a “random walk on a graph”, in the sense that $\mathbf{A}(i, j)$ represents the probability of “walking” in one step from \mathbf{x}_i to \mathbf{x}_j . One may show that the probability of such transitions for a t -step random walk is defined by the matrix product \mathbf{A}^t .

The semi-supervised classifier is defined as

$$p(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{N+M} \mathbf{A}^t(i, j)p(y_i = 1|\mathbf{x}_j, \mathbf{w}) \quad (11)$$

where $p(y_i = 1|\mathbf{x}_j, \mathbf{w})$ is defined as in (8). What (11) says is that the label associated with \mathbf{x}_i should be consistent with the labels associated with all other examples \mathbf{x}_j that are within a t -step random walk of \mathbf{x}_i . We again employ a prior on \mathbf{w} , $p(\mathbf{w})$. It is therefore important to note that the classification of any feature vector is *not* performed in isolation, rather, it is performed within the context of all unlabeled data from the site under test [3].

C. Active learning

When presenting the supervised and semi-supervised algorithms above, it was assumed that we had access to an *appropriate* set of labeled data $\{\mathbf{x}_n, y_n\}_{n=1, N}$. This assumption may not be valid in practice for UXO sensing, this motivating the idea of active learning, or *in situ* learning [4]. Specifically, for a supervised or semi-supervised algorithm we may use the Bayesian formalism to estimate a posterior density function on the model parameters \mathbf{w} , as expressed in (9). In active learning we approximate this posterior as a multivariate Gaussian in the parameters \mathbf{w} , from which we may compute the uncertainty in the model parameters based on the labeled data considered thus far (this defined by the covariance matrix). In active learning we sequentially ask which of the unlabeled signatures would be most informative (would most reduce the aforementioned covariance) if the associated labels could be acquired. The items associated with these most-informative signatures are excavated and the labels revealed, thereby providing a labeled set of data with which the classifier may be designed. The process is terminated when the information to be accrued based on new excavated labels is below a prescribed threshold.

With active learning, excavation is performed in two stages. First, assuming no *a priori* labeled data, items are excavated with the purpose of acquiring most-informative labels. After this phase is completed, a supervised or semi-supervised classifier is designed, in the manner discussed above. In the second phase of excavation, a prioritized dig list is provided based on classifier predictions.

V. Cost, Performance and Technology Limitations

A. Factors Affecting Cost and Performance

The cost of this Dem/Val (for SIG) is only the man hours required to perform data analysis, including feature extraction, algorithm design and testing. Delays pertaining to data collection and delivery were the only issues that may have affected the performance of the project; this turned out not to be an issue.

B. Advantages and Limitations of the Technology

An overarching problem in UXO remediation is the number of holes not associated with UXO that must be excavated to find all UXO in a given area. This false-alarm problem dramatically increases costs and slows remediation. Standard mag-and-flag techniques are particularly prone to the false alarm problem, since any ferrous subsurface anomaly is flagged for excavation. With the advent of digital geophysics, more automated techniques for anomaly selection have been developed. While some of these remain *ad hoc*, such as making declarations based on visually-derived features, others have begun to incorporate features associated with inversion of physics-based models into a statistical decision framework. The performance of visually-derived approaches is often defined by the experience of the analyst; current physics-based statistical approaches are limited by a lack of site-specific training data.

A significant benefit to the DoD of this project is the reduced number of total excavations required to clean a given site. Specifically, most of the false alarms associated with current technologies are attributed to the mismatch between available labeled training data and the unlabeled data to be classified. Active learning constitutes a state-of-the-art mathematically rigorous means of adaptively augmenting the training set, based on the site-dependent observed data. The active-learning framework provides two dig lists: the first defines a set of signatures for which access to the associated labels is most informative to classifier design, and after these labels are acquired a second dig list is provided specifically targeted toward excavating UXO.

Another key component of the technology to be demonstrated and evaluated is semi-supervised learning, in which classification of any given target is placed in the context of all targets of interest from a given site. As demonstrated in prior SERDP results, active and semi-supervised learning have resulted in significantly reduced total excavations, while achieving high UXO detection. Stated concisely, the techniques demonstrated here represent the state of the art in digital geophysics, and success on this project has the potential to transform the manner in which UXO cleanup is performed.

While the goal is to dramatically reduce the number of false alarms associated with a clean-up activity, it would be disingenuous to suggest that 100% of the UXO can be detected with no false alarms. There are certainly classes of clutter and geophysical anomalies that cannot be reliably differentiated from UXO, possibly as a result of data quality, or as a result of an inherent overlap of the two classes of objects in the feature space. While it is likely that the demonstrated techniques will reduce the false alarm problem, it is highly unlikely that they will mitigate the problem completely.

VI. Algorithmic Details for Sibert Data

A. Feature Selection and Threshold Settings

Within this study, Signal Innovations Group (SIG) performed feature extraction on the following data sets collected at the Sibert site: EM61, EM63, magnetometer and GEM3. The EM63 and GEM3 data are cued. The feature extraction was performed successfully on all data (although a portion of the EM61 and magnetometer data were not deemed to be of significant quality for feature extraction, with this discussed further below). For the magnetometer sensor, the standard dipole model was been employed [1], as discussed above, yielding two features: the dipole moment and the fractional error between the measured and modeled data (goodness of fit). For the EMI sensors we have employed the Duke-developed dipole model [5], also discussed above. The features from this model are two dipole moments, denoted $M1$ and $M2$, and with each of these moments is an associated “resonant” frequency, respectively $W1$ and $W2$ (the frequencies are actually imaginary, and $W1$ and $W2$ represent the associated magnitudes of these imaginary terms). The frequencies $W1$ and $W2$ correspond to decay constants in the time domain. Therefore, the models employed for the time and frequency domain EMI sensors have the same features. In addition to these four physics-based EMI features, there is a fifth EMI feature, corresponding to the fractional error (goodness of fit) between the modeled and measured data. In Figures 11-14 are shown the extracted features for the four sensors, where here we only show features for the labeled signatures (labels delivered by ESTCP). The features represented here give a sense of the degree of separation between the UXOs and non-UXOs (at least for the examples for which training data were provided). From these figures we note that, for the EM61, EM63 and MAG sensors, several of the two-feature combinations represented in Figures 11-14 yield good separation between UXO and non-UXOs. The cued GEM3 data appears to be of good quality, but the 5×5 grid employed in those measurements appears to be too sparse to yield a good inversion for the model parameters; this is deemed to be the principal reason for which the GEM3 features seem to show less separation between UXO and non-UXO. SIG considered classification on seven sets of features, in particular using features from the (*i*)

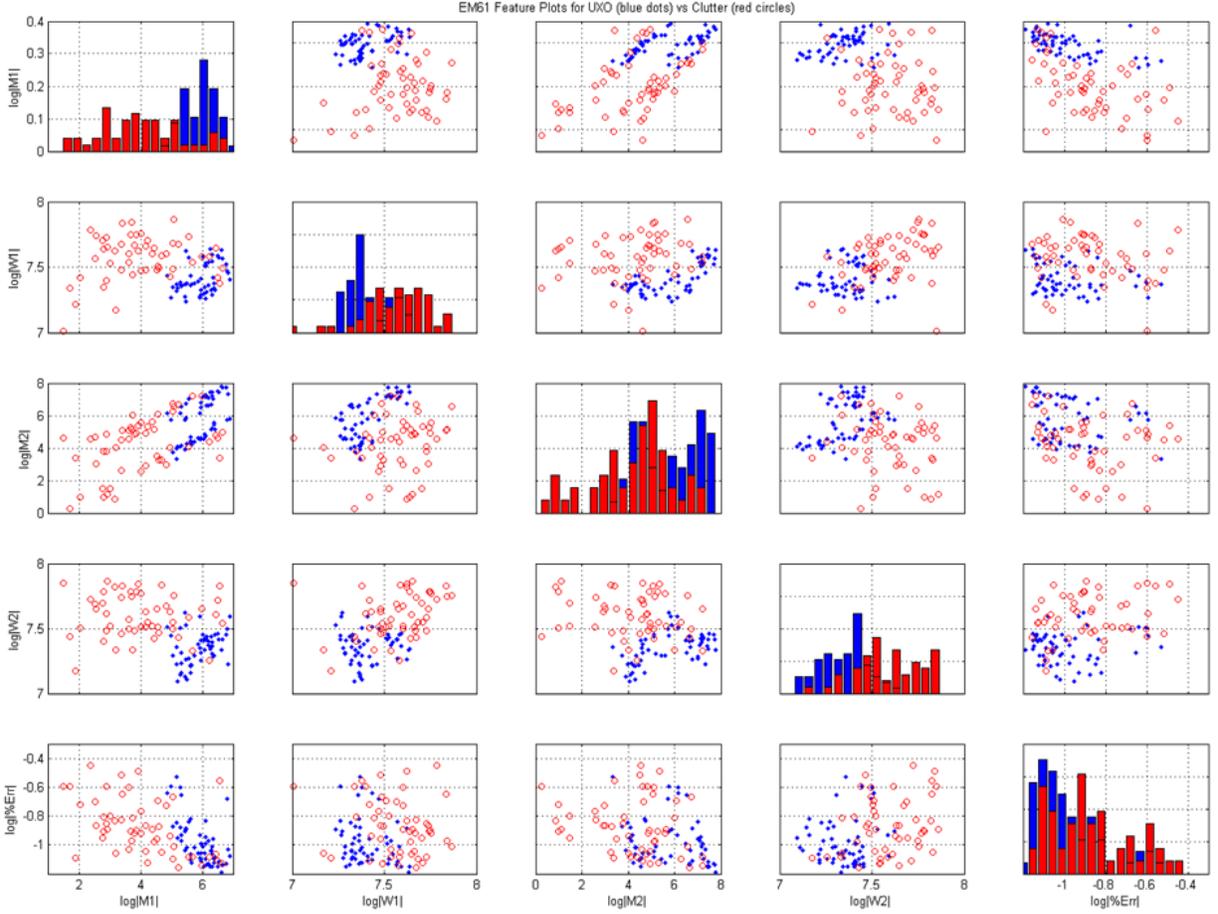


Fig. 11. Features extracted from the EM61 sensor, for *labeled* measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.

MAG sensor, (ii) EM61, (iii) EM63, (iv) GEM3, (v) MAG and EM61, (vi) MAG and EM63 and (vii) MAG and GEM3. For each of these seven types of feature combinations, we considered both a supervised [1], [6] and semi-supervised [7] classifier. In addition to performing classification using the given labeled data, we also considered active-learning [8].

B. Detailed Aspects of the Analysis

1) **Analysis decisions:** Based upon SIG’s experience with the Ft Ord EM61 data, we used all four time gates in the provided EM61 data (see the discussion in Section 3.3). Further, we have

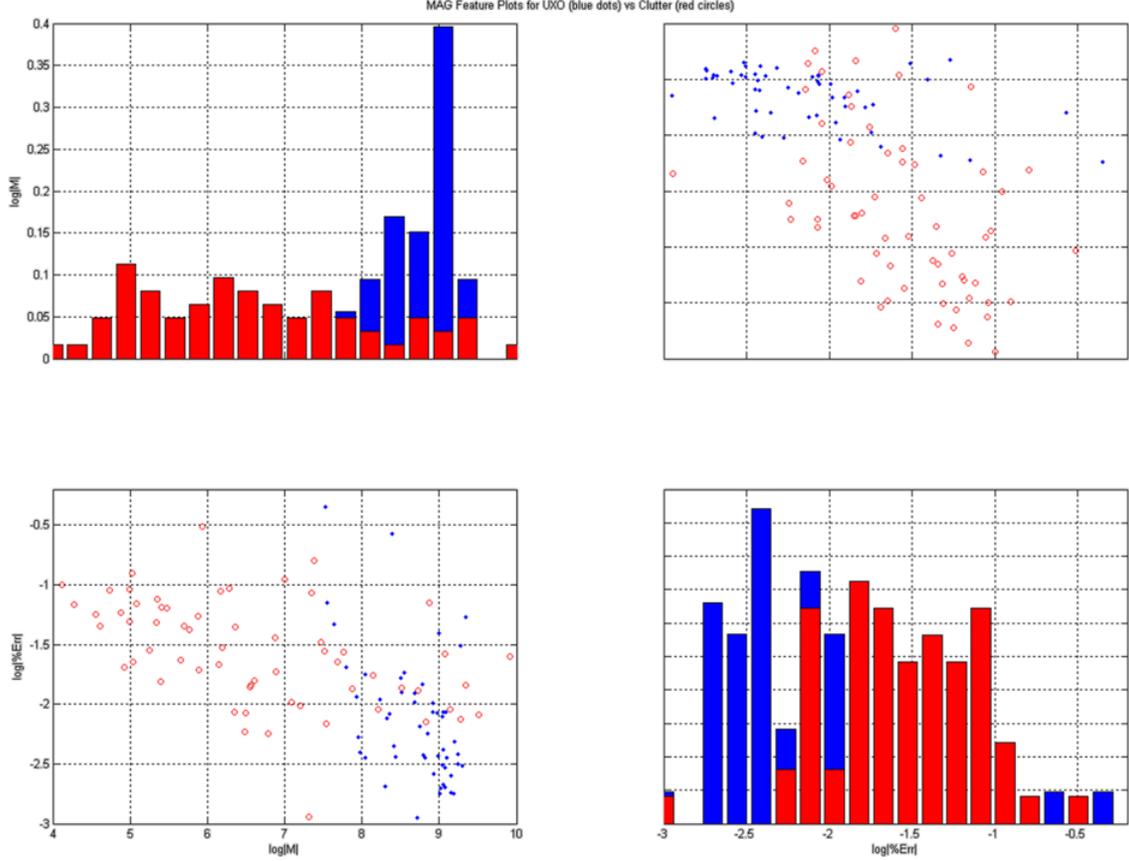


Fig. 12. Features extracted from the magnetometer sensor, for *labeled* measured at the Sibert site. The features are the dipole moment and model-fit error (from top to bottom, and left to right). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.

also employed all of the high SNR time gates provided with the EM63 data (see Section 3.2). We have employed all of the frequency-dependent, cued GEM3 data as given to us, although as indicated above we believe the 5×5 spatial grid associated with that data are insufficient for accurate estimation of the EMI model parameters (although we used these GEM3-derived target parameters as best as we could in the classification study).

2) **Parameters Estimated:** Above we have described the target parameters that have been estimated via the measured data. The same model parameters are estimated for both the time and frequency domain EMI sensors, although the estimations are performed in different ways. Although we also estimate the depth of the targets, with both the EMI and magnetometer sensors,

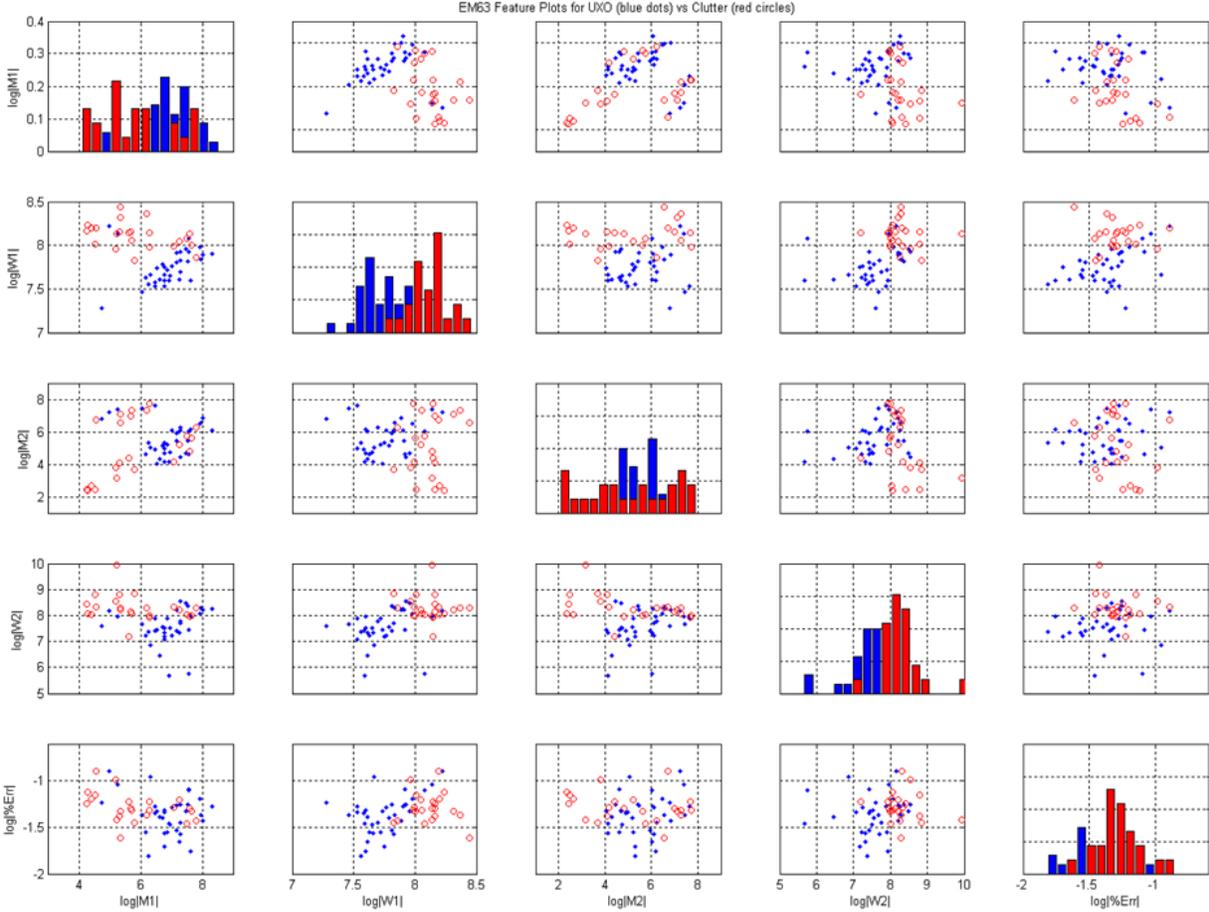


Fig. 13. Features extracted from the EM63 sensor, for *labeled* measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.

depth was not used as a feature within the classifier. We note from Figures 11-14 that, based on the labeled data, there appear to be some feature combinations that provide better separation than others. However, the classifiers sort this out on their own, without human intervention, and therefore we used all features (except depth) within our classifier; this was done for both the supervised and semi-supervised classifiers. Examples of how the supervised and semi-supervised algorithm select/weight features is discussed below in Section 6.3.

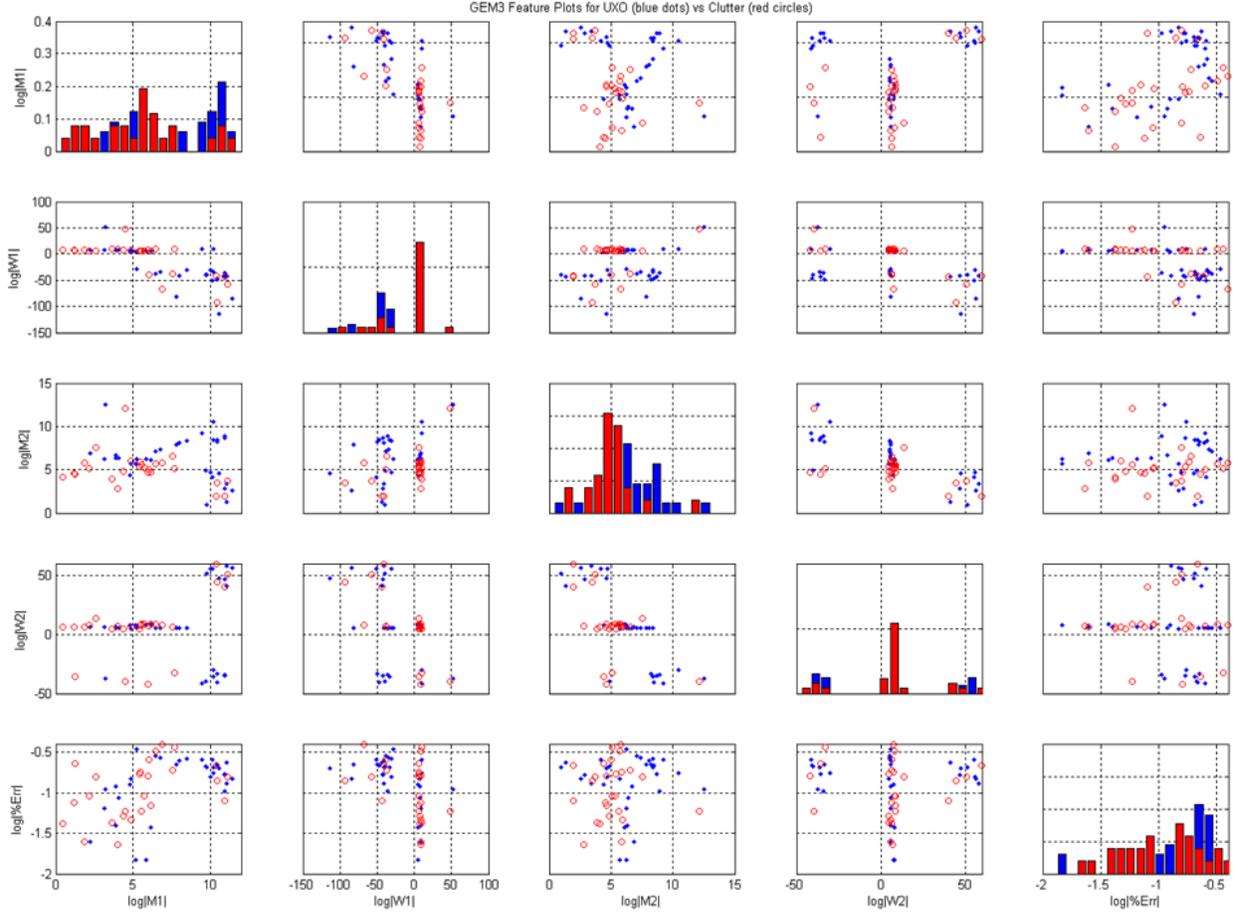


Fig. 14. Features extracted from the GEM3 sensor, for *labeled* measured at the Sibert site. The features are two dipole moments $M1$ and $M2$, and associated “resonant” frequencies $W1$ and $W2$ (the latter correspond to the respective decay constants in the time domain). The features are ordered, from top to bottom (and left to right), $M1$, $W1$, $M2$, $W2$, and Err , and the log of each feature is plotted, as this is what is used in the final classifier; the fifth (last) feature is the goodness of fit (model error relative to the measured data). The off-diagonal plots show all combinations of viewing two features at a time. Along the diagonal, a histogram is shown for the distribution of each individual feature, with the UXO and non-UXO histograms depicted in different colors. Blue: UXO, Red: non-UXO.

3) **Setting Thresholds:** The classifiers we employ yield a statistical estimate of the label of a given item [1], [7]. Specifically, let \mathbf{x} represent a given feature vector under test, and our goal is to estimate the label l , where $l = 1$ is chosen to correspond to a UXO and $l = 0$ to a non-UXO. Our algorithms yield the probability $p(l = 1|\mathbf{x})$, and $p(l = 0|\mathbf{x}) = 1 - p(l = 1|\mathbf{x})$. Let the cost of declaring an item to be a UXO when it is actually a non-UXO be denoted C_{10} , while the cost of declaring an item non-UXO when it is actually a UXO is denoted C_{01} . We set the cost (reward) associated with making a correct classification to zero: $C_{11} = C_{00} = 0$. Given a feature

vector \mathbf{x} under test, the *expected* cost (or risk) of declaring the associated item to be a UXO is

$$R_{UXO} = C_{10}p(l = 0|\mathbf{x}) \quad (12)$$

while the risk of declaring the item to be non-UXO is

$$R_{non-UXO} = C_{01}p(l = 1|\mathbf{x}) \quad (13)$$

Our objective is to minimize the risk, and therefore we declare \mathbf{x} to be a UXO if $R_{UXO} < R_{non-UXO}$, and otherwise we declare non-UXO. Hence, from (12) and (13), we declare the item to be a *non-UXO* if $\frac{p(l=0|\mathbf{x})}{p(l=1|\mathbf{x})} > \frac{C_{01}}{C_{10}}$. Thus, by selecting the costs C_{01} and C_{10} , and given a statistical measure $p(l = 1|\mathbf{x})$, one defines the threshold or operating point on the ROC. Note that the more the relative cost of a missed UXO increases, corresponding to increasing $\frac{C_{01}}{C_{10}}$, the greater the ratio $\frac{p(l=0|\mathbf{x})}{p(l=1|\mathbf{x})}$ required to leave an item unexcavated (*i.e.*, the more confident one must be in the declaration of a non-UXO).

From the above discussion, the absolute values of C_{01} and C_{10} are unimportant, rather the ratio $\frac{C_{01}}{C_{10}}$ defines the threshold (the ratio $\frac{C_{01}}{C_{10}}$ represents how more costly it is to leave a UXO unexcavated, relative to the cost of a false alarm). Therefore, in setting our threshold, we will set this ratio. As one varies this threshold, one maps out the receiver operating characteristic (ROC).

The question then reduces to: how does one set the ratio $C = C_{01}/C_{10}$? In our previous Ft Ord studies we set $C = 100$, which implies that the cost of a missed UXO is 100 times more costly than a false alarm. Based on a leave-one-out analysis of the labeled data, we assessed how best to select the threshold C for defining our dig lists. This analysis is detailed below in Section 7.

4) Analysis of GPO Data: In Figures 11-14 are shown the distribution of the features for the labeled UXO and non-UXO. Most of the labeled UXO come from the GPO region, and therefore these plots provide a representation of how variable the target (UXO) parameters are. Based on our initial analysis, the EM61, EM63 and magnetometer data seem to be yielding reliable features. Our analysis indicates relatively poor fits to the measured GEM3 data via the

aforementioned EMI model, which we attribute to not enough spatial samples in the cued GEM3 data.

C. Feature selection/weighting

Let \mathbf{x} represent a feature vector under test. The supervised and semi-supervised algorithms seek to quantify the probability that \mathbf{x} is associated with a UXO:

$$p(l = 1|\mathbf{x}, \theta) = \sigma(\mathbf{x}^T \theta) \quad (14)$$

where superscript T represents vector transpose and

$$\sigma(y) = \exp(y)/[1 + \exp(y)] \quad (15)$$

Note that the logistic link function $\sigma(y)$ yields a probability that is bounded between zero and one, approaching one as y becomes large and positive, and approaching zero as y becomes large and negative. The vector θ weights each of the components that define \mathbf{x} , implying it weights the importance of the feature components in \mathbf{x} ; we append a one to the vector \mathbf{x} , this representing a “bias” term [6]. A “shrinkage” or “sparseness” prior is usually employed on θ [6], which encourages many of the components of θ to be small, and in this sense the θ serves to select the important features, and de-emphasize the unimportant features. This is done automatically by the algorithm, and hence there is no need for human selection of features. This same analysis is performed on all seven sensor combinations, and for both the supervised and semi-supervised algorithms.

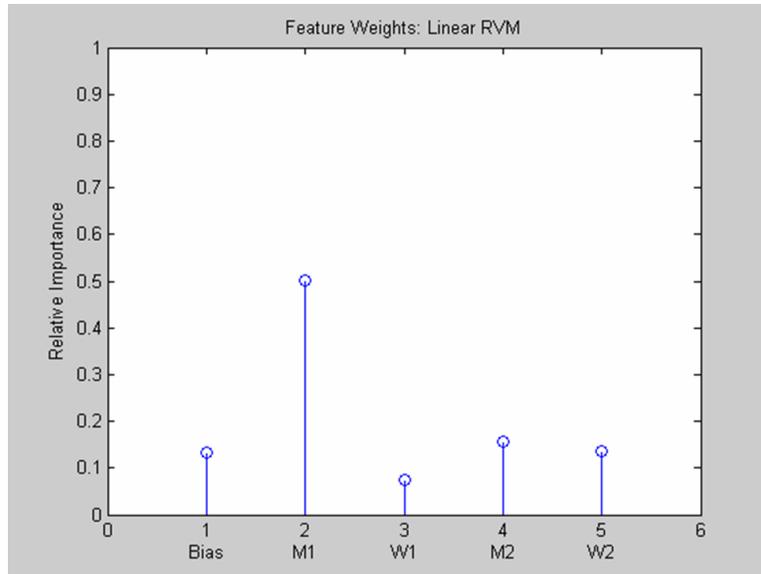


Fig. 15. Weights on the vector θ as computed for the EM61 sensor, using a supervised classifier.

To demonstrate this feature weighting, we consider as an example supervised processing of the EM61 data, with the learned weights depicted in Figure 15. It is demonstrated in this figure that all of the features are relatively useful, although the first dipole moment $M1$ is the most important feature. This is expected from Figure 11, which shows good separation in this feature, between UXO and non-UXO. It is important to note that we do not explicitly remove any features from the analysis, rather the algorithm determines the relative importance of the features.

VII. Details on Setting Thresholds

A. Supervised vs semi-supervised learning and classification confidence

SIG provided dig lists for seven different sensor combinations (EM61, Mag, EM63, GEM3-cued, plus combining each of the EMI sensors with Mag). For each of these seven combinations, we provided dig lists based on a supervised and semi-supervised classifier. There are several questions that should be addressed in this context, principally how to set the threshold and how to assess confidence. We assess the latter question first.

Concerning assessing confidence in the classification decision, note that both the supervised and semi-supervised classifiers [6], [7] yield explicit probabilistic measures as an output. Specifically, given a feature vector \mathbf{x} under test, both the supervised and semi-supervised algorithms yield probabilistic outputs $p(l = 1|\mathbf{x})$, where the label $l = 1$ corresponds to a UXO, and the label $l = 0$ corresponds to a non-UXO; the probability of a non-UXO is $p(l = 0|\mathbf{x}) = 1 - p(l = 1|\mathbf{x})$. Therefore, the supervised and semi-supervised algorithms explicitly give a probabilistic measure in the confidence of declaring an item a UXO: the higher the probability $p(l = 1|\mathbf{x})$, the more confident the algorithm is that the item under test is a UXO.

The matter of classifier confidence and supervised versus semi-supervised learning is of significant importance, and therefore a further discussion is provided here. Recall the key distinction between a supervised and semi-supervised classifier: A supervised classifier is based only on the labeled data (data for which we have feature vectors *and* labels, this often termed the “training” data), while a semi-supervised classifier is designed based on the labeled and unlabeled data (the unlabeled data corresponds to feature vectors for which we only have feature vectors, with this often termed the “testing data”). A semi-supervised algorithm employs the context provided by the unlabeled data when learning the classifier parameters, while a supervised classifier only employs the labeled data (doesn’t exploit context). We here discuss how this impacts the classification decisions and confidence levels in the study.

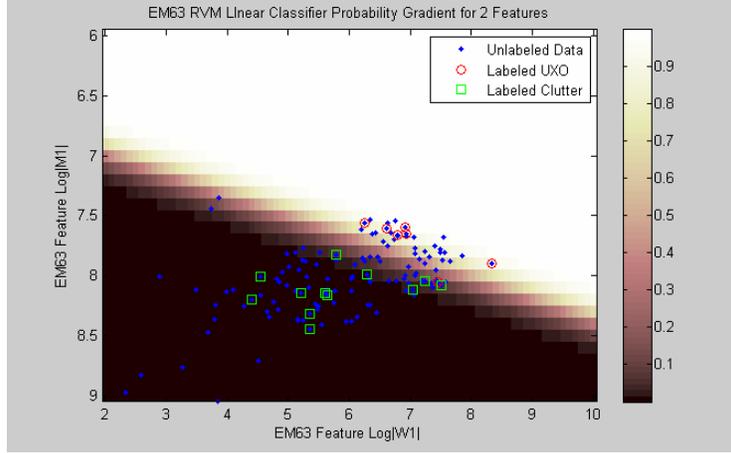


Fig. 16. The probability of being a UXO, $p(l = 1|\mathbf{x})$ is plotted in a two-dimensional feature space characteristic of two of the features in the overall feature vector \mathbf{x} . Results are shown for the EM63 sensor, using the labeled and unlabeled data from the Sibert site. In this plot a supervised classifier is considered, and therefore $p(l = 1|\mathbf{x})$ is designed using only the labeled data. An important thing to note is how confident the classifier is in the decision boundary: above the boundary for which $p(l = 1|\mathbf{x}) = 0.5$ one observes that $p(l = 1|\mathbf{x}) \approx 1$ very quickly, and below $p(l = 1|\mathbf{x}) = 0.5$ we observe $p(l = 1|\mathbf{x},) \approx 0$ very quickly as a function of \mathbf{x} . Hence, based on the limited available labeled data, and in absence of the context provided by the unlabeled data, the classifier is very confident in what parts of feature space \mathbf{x} correspond to UXO and non-UXO. The classifier is designed based on all EM63 features, and the plot here considers the classifier decision as viewed in a two-dimensional plane within that feature space (corresponding to the decay constant and moment associated with the first EMI dipole).

To focus the discussion, Figures 16 and 17 show, respectively, the classifiers learned using a supervised and semi-supervised classifier, for the EM63 data at the Sibert site. There are two important distinctions between the learned classifiers $p(l = 1|\mathbf{x})$ under the supervised and semi-supervised conditions: (i) there is a slight shift in the boundary for which $p(l = 1|\mathbf{x}) = 0.5$, and (ii) $p(l = 1|\mathbf{x}) \approx 0.5$ for a much larger portion of the feature space \mathbf{x} in the semi-supervised case (Figure 17) as compared to the supervised case (Figure 16). The latter point is particularly important in the context of defining a threshold for the classifier, because the context provided by the unlabeled data is causing the semi-supervised classifier to be less certain in which features \mathbf{x} correspond to UXO.

Recall from the discussion above that an item is declared to be non-UXO if $\frac{p(l=0|\mathbf{x})}{p(l=1|\mathbf{x})} > C$. For a selected value of the constant C , and using $p(l = 1|\mathbf{x}, \theta) = 1 - p(l = 0|\mathbf{x})$, this expression places the requirement that a feature vector is deemed associated with non-UXO if $p(l = 0|\mathbf{x}) > C/(1+C)$. Therefore, for large C one must be very confident that an item is not a UXO, *i.e.*, we require $p(l = 1|\mathbf{x}) \approx 0$ to leave an item unexcavated. In Figure 16 we note that for the supervised

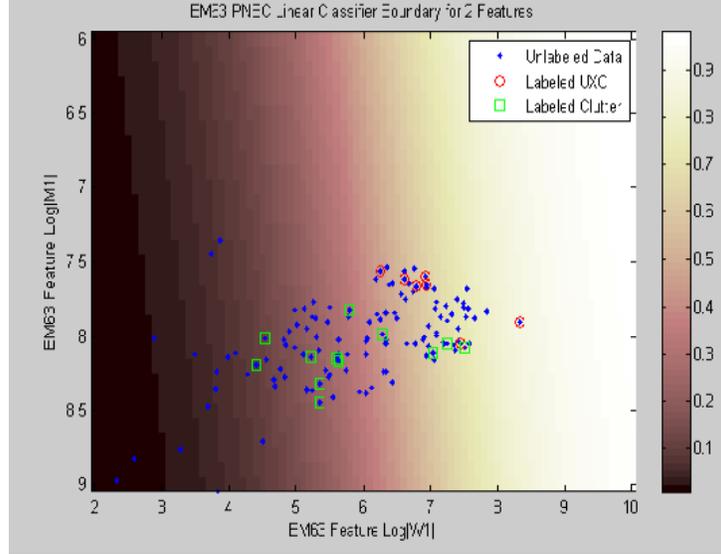


Fig. 17. The probability of being a UXO, $p(l = 1|\mathbf{x})$ is plotted in a two-dimensional feature space characteristic of two of the features in the overall feature vector \mathbf{x} . Results are shown for the EM63 sensor, using the labeled and unlabeled data from the Sibert site. In this plot a semi-supervised classifier is considered, and therefore $p(l = 1|\mathbf{x})$ is designed using both the labeled and unlabeled data. In comparison to Figure 16 note that the $p(l = 1|\mathbf{x}) = 0.5$ boundary is shifted slightly; of more importance, note the far more gradual change in the probabilities $p(l = 1|\mathbf{x})$ for features \mathbf{x} away from the region $p(l = 1|\mathbf{x}) = 0.5$. This implies that the semi-supervised classifier is less confident in which features \mathbf{x} correspond to UXO, as a result of the context provided by the unlabeled data. The classifier is designed based on all EM63 features, and the plot here considers the classifier decision as viewed in a two-dimensional plane within that feature space (corresponding to the decay constant and moment associated with the first EMI dipole).

classifier, for which the contextual information from the unlabeled data is *not* used, the space of feature vectors \mathbf{x} for which $p(l = 1|\mathbf{x}) \approx 0$ is relatively large. By contrast, by including the context provided by the unlabeled data, the semi-supervised classifier is far more conservative, since the region of feature vectors for which $p(l = 1|\mathbf{x}) \approx 0$ is much reduced. Therefore, we emphasize the *key distinction between the supervised and semi-supervised algorithms*: It is anticipated that the semi-supervised algorithm will be far more conservative than its supervised counterpart, and therefore that the semi-supervised algorithm will excavate more of the site than the supervised algorithm. This expectation is demonstrated below when presenting supervised and semi-supervised analysis of the Sibert data, across all seven sensor combinations.

B. Leave-one-out analysis of Sibert data

To examine setting the threshold on the supervised and semi-supervised classifiers, we consider leave-one-out training. For the supervised algorithm as applied to the seven data sets, the ROC curve was generated as follows. Assume that there are N labeled examples. We take $N - 1$ of these and design a classifier, and then apply it to yield the probability $p(l = 1|\mathbf{x})$ for the N th feature vector \mathbf{x} which was held out. This is done for all N hold-out combinations, from which $p(l = 1|\mathbf{x})$ is computed for all N labeled examples. By running a threshold across $p(l = 1|\mathbf{x})$, a ROC curve is generated. For the semi-supervised algorithm, the same process is performed, but now the unlabeled data are also employed when learning each of the N classifiers. In practice the classifier so learned is far more stable/repeatable among the N different hold-out cases for the semi-supervised algorithm as compared to the supervised classifier, since the semi-supervised algorithm is always using the same (large) set of unlabeled data. While the ROCs so learned are useful, of far more importance is selection of the threshold, which in the context of the discussion above implies selection of the constant C .

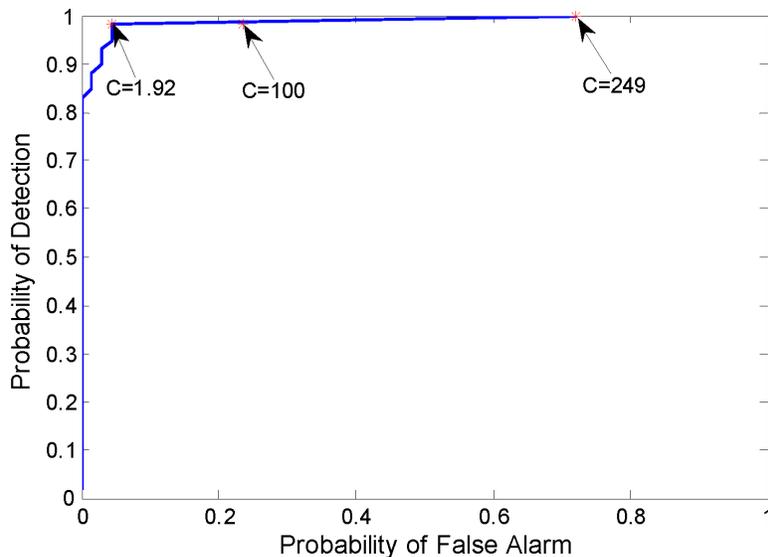


Fig. 18. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM61 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

We first consider the EM61 data, where in Figures 18 and 19 we plot the ROCs for, respectively, supervised and semi-supervised learning. While the ROCs are useful in some sense, of far more importance is to examine setting of the threshold C . In Figures 18 and 19 we identify different choices for the threshold. As was suggested by the analysis in Section 7.1, note that if one considers a fixed factor C , implying a fixed risk level, the semi-supervised classifier will be far more conservative than the supervised classifier (note that for $C = 100$, the semi-supervised classifier operates at a much higher false alarm rate than the supervised classifier). This is an important indication that by exploiting the context associated with the unlabeled data, *the semi-supervised algorithm is much more cautious in deciding which items to leave unexcavated*. With ESTCP's consent, to explore this issue, which we find interesting, and potentially an important distinction between supervised and semi-supervised learning, we will fix the constant C for the supervised and semi-supervised classifiers, and therefore the semi-supervised classifier will leave less items unexcavated (but also hopefully have a higher detection rate). Note that the ROCs are also fairly different, in that the supervised leave-one-out curve looks much better than the semi-supervised one. However, we anticipate that the semi-supervised one will be more indicative of what will occur with the unlabeled data, since it exploits the information associated with the unlabeled data. By contrast, the ROC in Figure 18 is only trained using the relatively small quantity of labeled data (over-training is likely).

In Figures 20 and 21 are shown, respectively, supervised and semi-supervised analysis of the magnetometer data. Note that in this case the ROCs are similar, which suggests that the labeled data may be a better representation of the unlabeled data. The most important thing to note is the different false-alarm rate one operates in a supervised and semi-supervised mode for a fixed cost ratio of $C = 100$. This strongly underscores the more-conservative nature of the semi-supervised classifier, as discussed in Section 7.1. Again, with ESTCP consent, and to explore this issue, when providing our dig list we will employ a fixed cost C for the supervised and semi-supervised classifiers.

In the following we present ROCs for supervised and semi-supervised analysis of the remaining sensor combinations, and note operating points (choices of C) in each of these. At the end of this discussion we discuss selecting the cost C in the Sibert study, as applied to the unlabeled

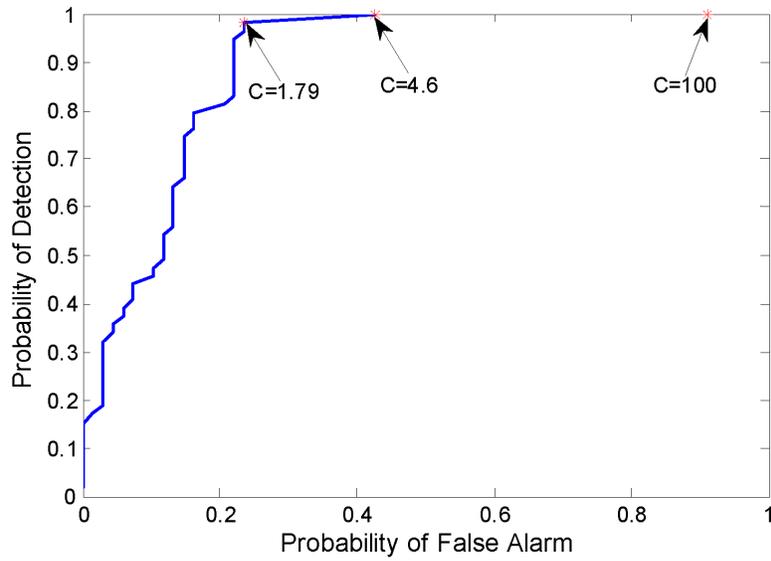


Fig. 19. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM61 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

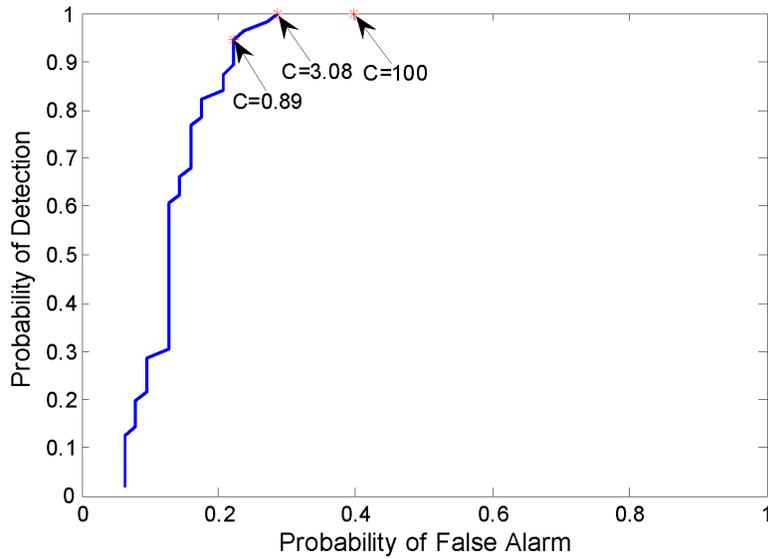


Fig. 20. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

data.

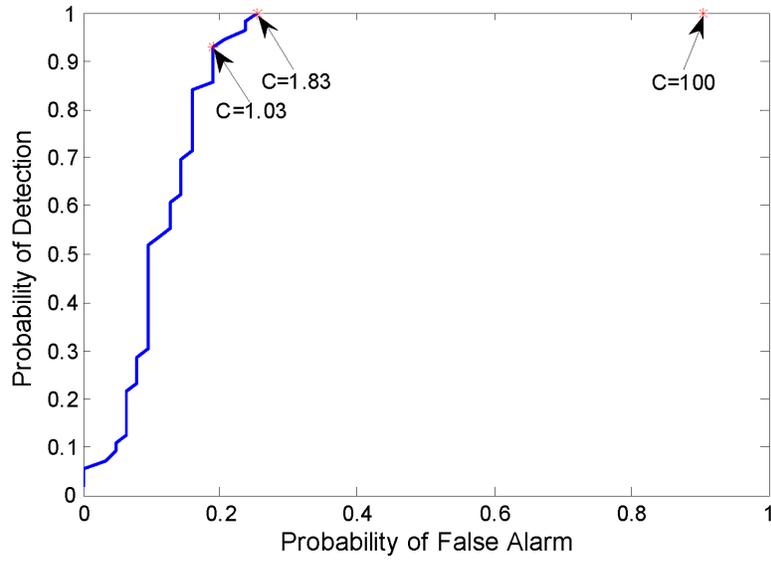


Fig. 21. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

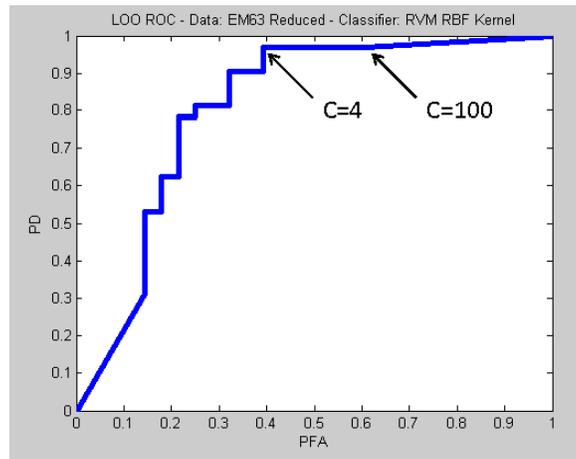


Fig. 22. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM63 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

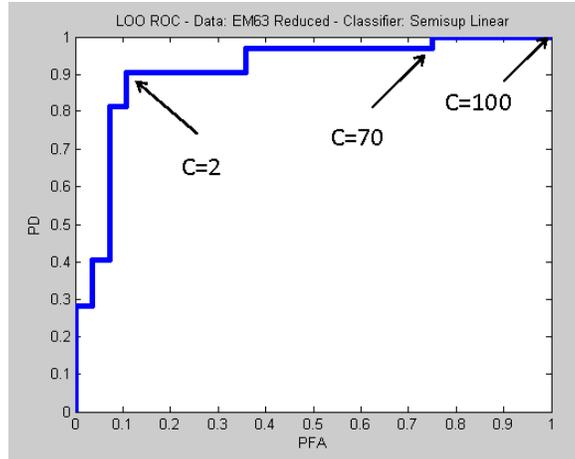


Fig. 23. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled EM63 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

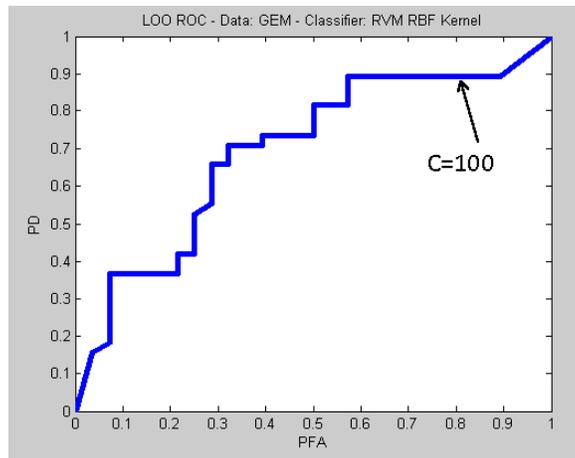


Fig. 24. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled GEM3 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

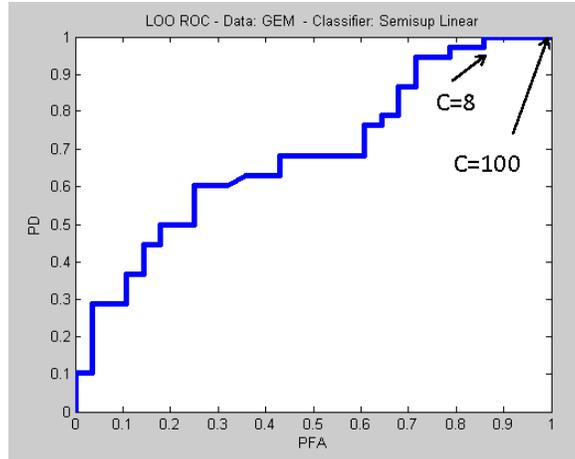


Fig. 25. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled GEM3 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

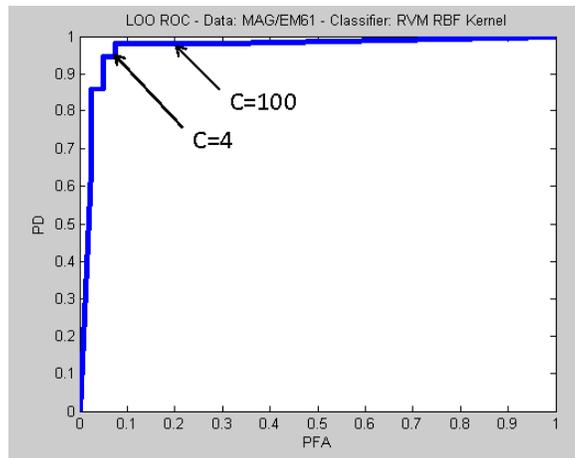


Fig. 26. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and EM61 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

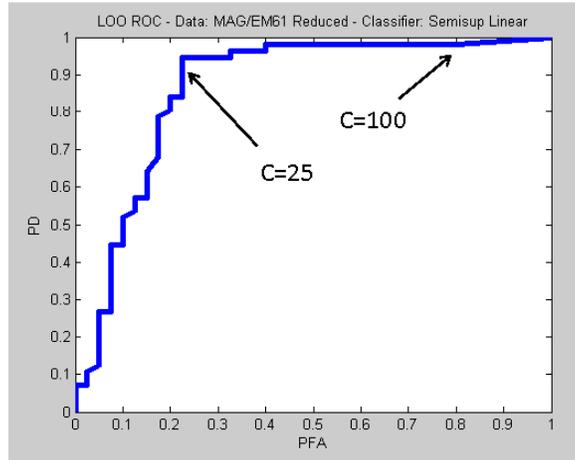


Fig. 27. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and EM61 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

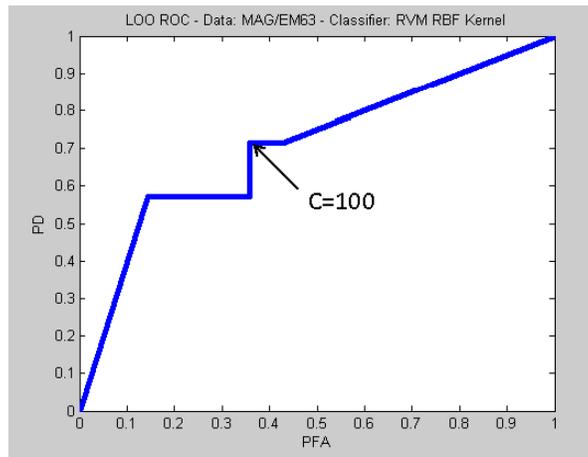


Fig. 28. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and EM63 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

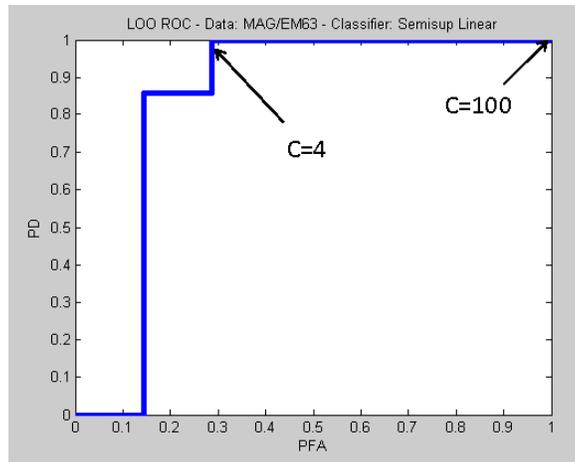


Fig. 29. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and EM63 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

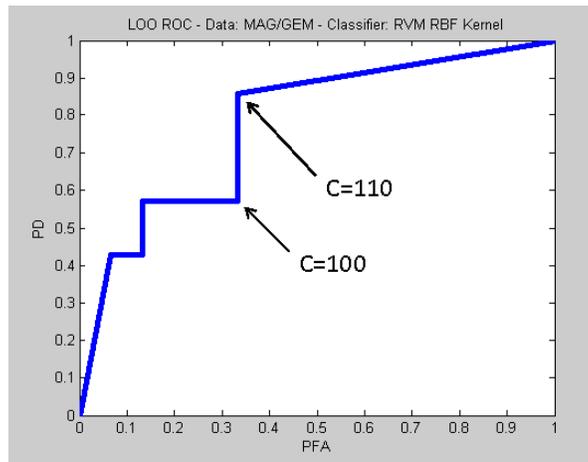


Fig. 30. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled combined magnetometer and GEM3 data from the Sibert site. Results are shown for supervised learning, and several different operating points, or thresholds C , are depicted.

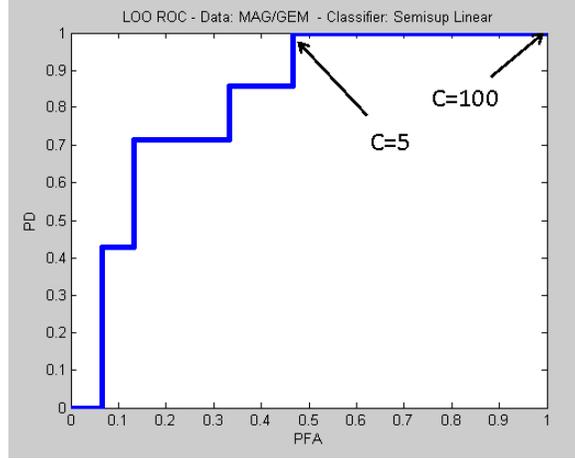


Fig. 31. Receiver operating characteristic (ROC) for leave-one-out analysis of the labeled magnetometer and GEM3 data from the Sibert site. Results are shown for semi-supervised learning, and several different operating points, or thresholds C , are depicted.

C. Setting of thresholds for blind test

To define which items to leave unexcavated, we must define the cost ratio C . To give a sense of what different values of C mean, we recall that an item is declared non-UXO and left unexcavated if $p(l = 0|\mathbf{x}) > C/(1 + C)$. If we consider $C = 100$, an item is unexcavated if we are 99% sure it is non-UXO. By contrast, setting $C = 50$ implies an item is left unexcavated if we are 98% sure it is non-UXO. Finally, setting $C = 25$ implies we leave an item unexcavated if we are 96% sure it is non-UXO.

From the leave-one-out analysis, it is clear that relatively high-confidence declarations are possible based on the EM61 and magnetometer data, and to a lesser extent with the EM63 data. Also, consistent with the discussion in Section 7.1, the semi-supervised algorithm will be less confident (more conservative) than the supervised algorithm, and will leave fewer items unexcavated.

VIII. Items Excluded from Classification Study

A. Subjective data removal

Within the context of this study, SIG took the perspective that our principal objective was to leave no UXO unexcavated. Therefore, this implies that we sought to excavate any item for which we had any doubt in our ability to perform classification. In fact, our main objective was to perform classification on signatures that were deemed of high-enough quality for this purpose, and therefore we were very conservative in defining which items to perform classification on (in hindsight, we were probably *overly* conservative). In the pre-processing step, SIG determined which items were deemed to have signatures of sufficient quality for subsequent classification. All of the signatures were examined via *visual inspection*. Further, the data from each item of sufficient signature quality were submitted to the dipole models discussed above. Using the dipole-model parameters extracted from the data, the model may be used to reconstitute the data, and a goodness of fit (GOF) could be computed (Euclidian distance between the measured data and the associated modeled data). If the GOF was deemed to be of poor quality relative to the average across all items, then the item associated with the poor-fit data were excluded from the classification study. In addition, recall that we must perform a nonlinear fit of the model to the data. This fit may be susceptible to the initialization of the gradient-search method. If the features were unreliable, in the sense that they were very sensitive to the initialization, the associated data were not considered further for classification. A summary of the number of total items per sensor and the associated number of excluded items is presented in Figure 32. Note that we also excluded data from the training set, so that the number of items used for classifier training was smaller than the number given.

It is important to recognize that the dipole models are designed to fit UXO and UXO-like clutter. If an item is very distinct from a UXO, it will not fit the model well. SIG recognized prior to the list submission that most if not all of the excluded items were not UXO, based on the poor fit to the model (and also because *all of the similarly removed training data were also not UXO*). However, in many of these cases the fit was so poor as to not be trusted at

	EM61	Mag	EM63	GEM3
Total Number of Items	908	1007	216	216
Number of Labeled Items	174	182	66	66
Total Number Excluded	344	396	22	0
Number of Labeled Items Excluded	49	63	8	0

Fig. 32. A summary of the number of signatures per sensor, the number of labeled training examples given per sensor, the total number of excluded items, and the number of these that came from the training set. None of the items excluded from the EM61 and magnetometer sensors was UXO.

all, and therefore it was deemed inappropriate to submit these poor/unreliable features to the classifier and attempt classification. Motivated as discussed above by the goal of leaving no UXO unexcavated, we simply did not make a classification decision on these excluded items (although we knew *a priori* that most if not all were not UXO). This point should be noted when considering the classification results below, which for the items on which classification *was* performed the performance is very encouraging. In some sense we didn't make classification decisions on the easiest items, those being ones for which the data didn't fit the UXO-based model at all. In retrospect, if we had declared as non-UXO all items that were significantly misfit by the model, or for which the model fit was unreliable, our performance would have been much better (we would have missed no more UXO, and the false alarm rate would have been much lower); we did not do this because, based on discussions with the sponsor, we were under the impression that classification decisions should only be made by the classifier, and not other considerations, as indicated above.

In Figures 33 and 34 we present examples of items that were excluded based on visual inspection and based upon the quality and reliability of the dipole-model fit to the data. Based on these data it was highly unlikely that these items were associated with UXO. However, since we could not explicitly run them through the classifier (the features were not reliable), we did not make a classification decision. By contrast, in Figure 35 we present example EM61 signatures

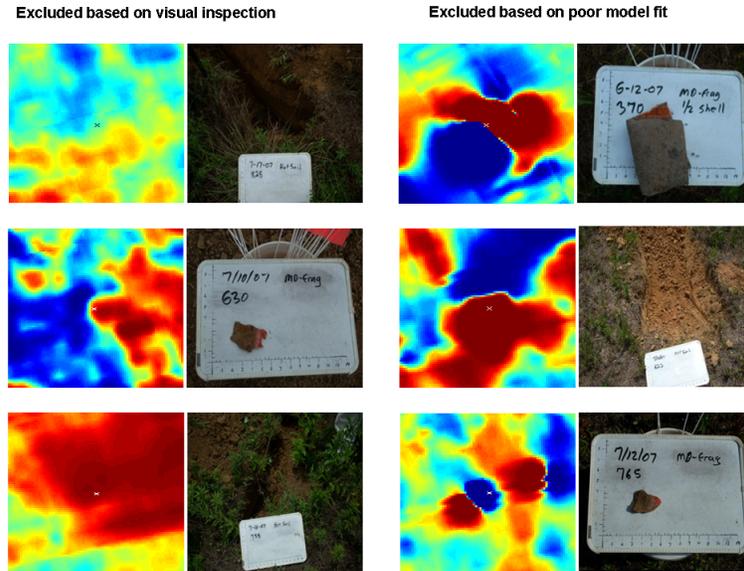


Fig. 33. Representative example data that were excluded from the study from the magnetometer sensor.

of high enough quality, in the sense that the associated dipole-model parameters were reliable and therefore appropriate for submission to a classifier.

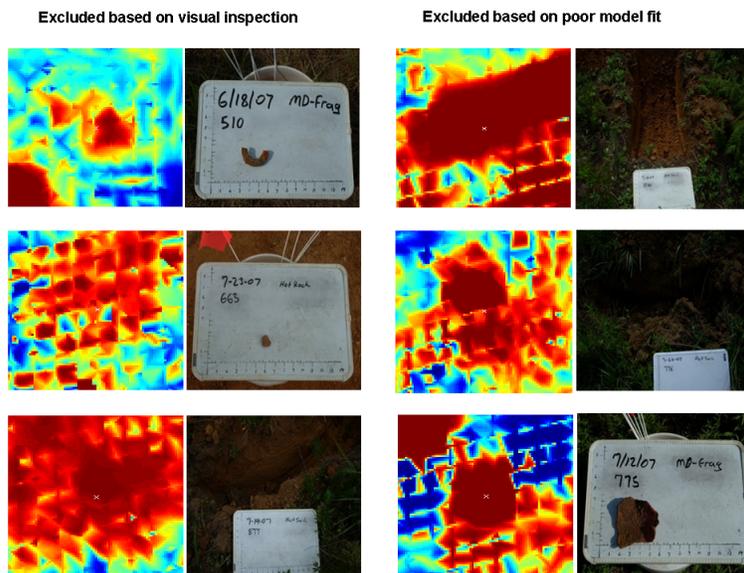


Fig. 34. Representative example data that were excluded from the study from the EM61 sensor.

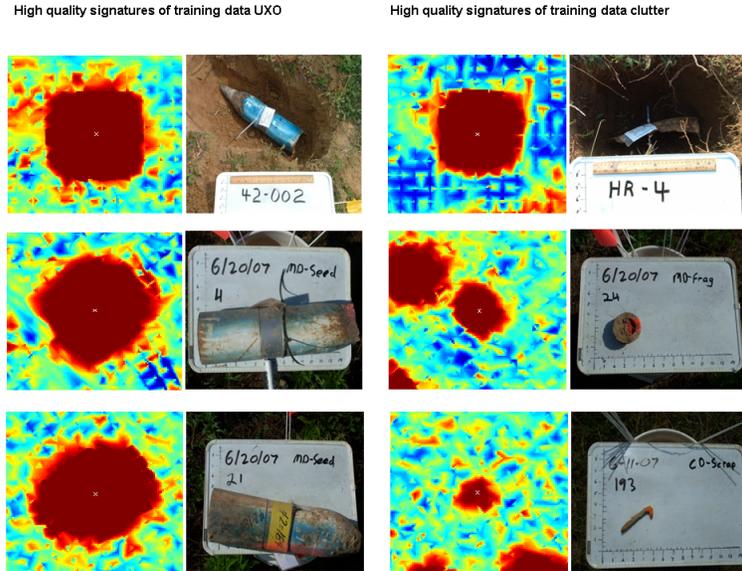


Fig. 35. Example EM61 signatures that were deemed of high-enough quality to perform classification

In Figure 8 we provide an example poor model fit for the EM63 sensor, with this an example of a target that was excluded from the classification analysis. In Figure 36 we present representative example EM61 measured data and the associated model reconstruction. All poor model fits, as in the left of this figure, were removed from classification analysis. We emphasize that there is certainly a subjective element to data removal (for classification), based on human analysis of data and model-fit quality.

B. Details on removal methodology based on data inspection

In Figure 36 we provided an example of a particular EM61 signature that was removed due to relatively poor data inversion. In addition, prior to the inversion, visual inspection was performed on every signature. In this phase a signature could be removed from subsequent analysis for one of three reasons:

- **Overlapping signals:** We did not perform inversion for cases in which multiple anomalies appeared to generate overlapping signatures. In principle one may perform inversion for overlapping signatures, but this is complicated, and we wished to avoid that in this study.
- **Anomaly confusion:** In this case there appeared to be multiple anomalies in the same

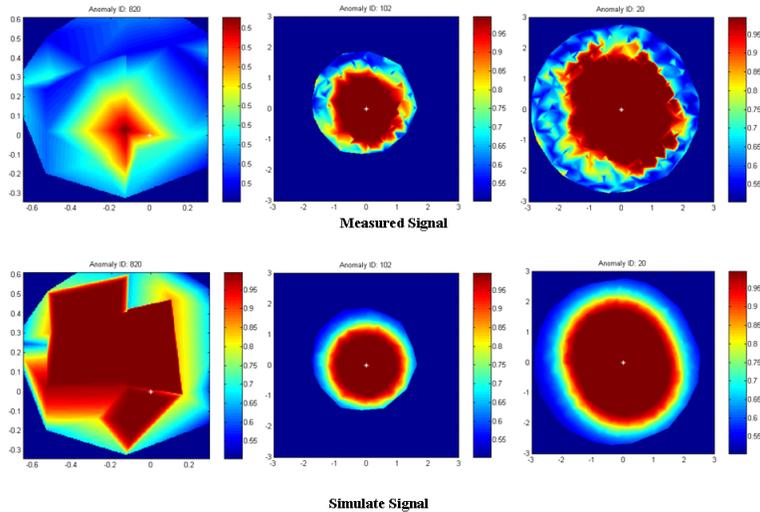


Fig. 36. Example measured EM61 signatures (top) and associated model reconstruction (bottom). The left-most example was removed from the classification analysis, and the middle and right data were used within the classifier. The white point in each figure is the as-given target location.

general location, and we could not unambiguously identify the signature associated with the corresponding target.

- **Weak signal:** In this case the signal had very weak amplitude, and therefore noise in the signature undermined inversion quality. This case may also be related to inaccuracies in the target location, because often while the signature directly at the specified target location was weak, there was a nearby (or multiple nearby) signature(s) with strong amplitude.

We provide examples of each of these three cases, for the EM61 and magnetometer sensors (Figures 37-42).

EM61 Overlapping Signatures

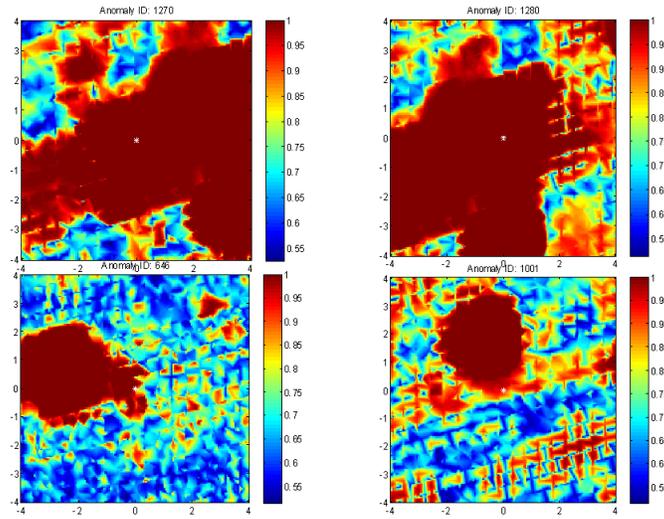


Fig. 37. Example measured EM61 signatures that were removed because of overlapping signatures. Note that the bottom two correspond to very-large signatures near the given target location (white dot), and we could not be sure if this was a location error or a strong nearby signature. This class corresponded to 30% of the removed cases.

EM61 – Anomaly Confusion

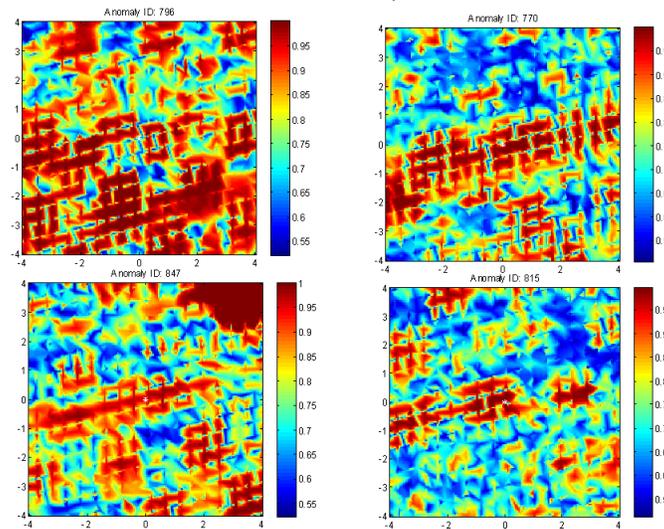


Fig. 38. Example measured EM61 signatures that were removed because it was confusing as to what precisely was the signature (highly anomalous signatures). This class corresponded to 55% of the removed cases.

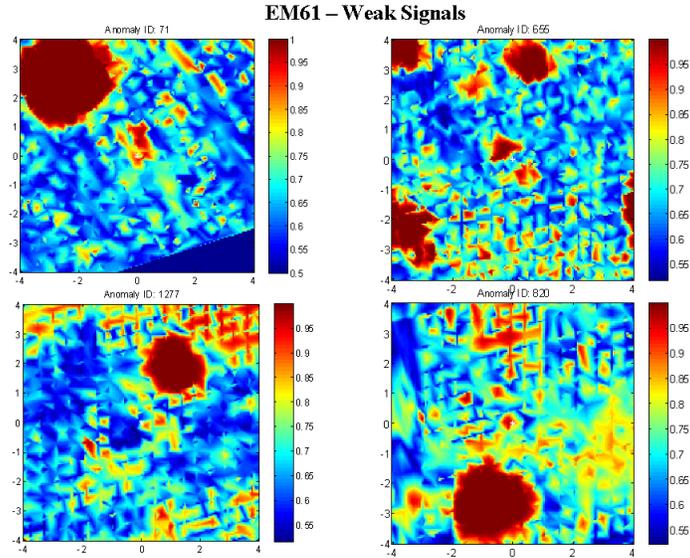


Fig. 39. Example measured EM61 signatures that were removed because of weak signals in the location of the specified target location (white point). Note that there are sometimes strong nearby targets and it is not clear if there is actually target-location error. This class corresponded to 15% of the removed cases.

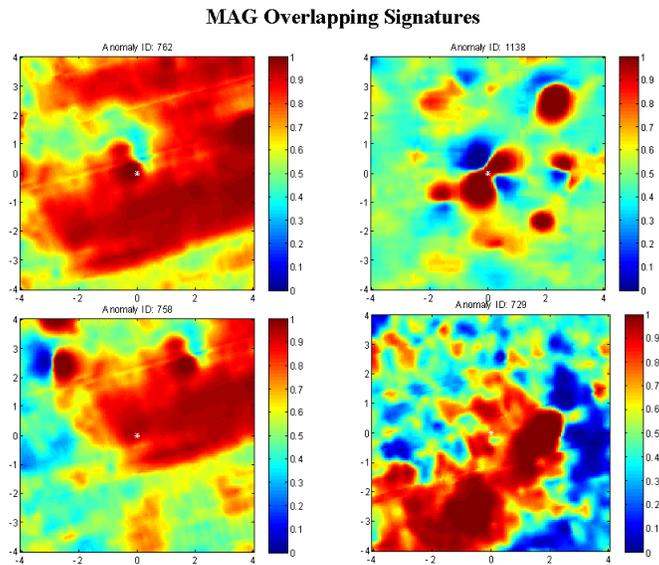


Fig. 40. Example measured magnetometer signatures that were removed because of overlapping signatures. Note that the bottom two correspond to very-large signatures near the given target location (white dot), and we could not be sure if this was a location error or a strong nearby signature. This class corresponded to 39% of the removed cases.

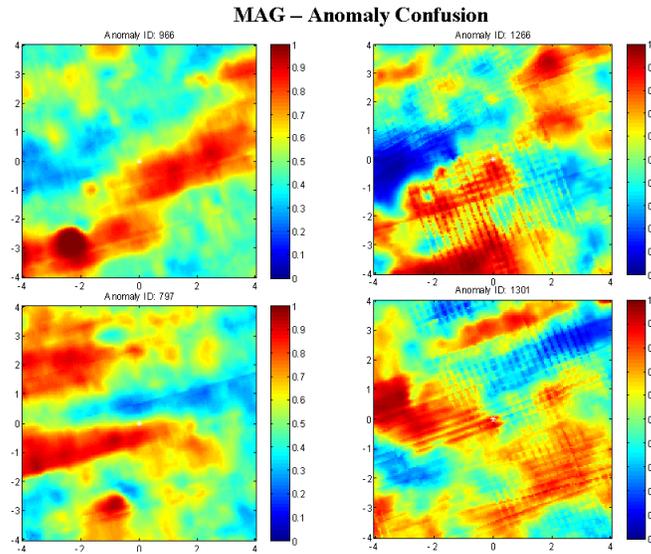


Fig. 41. Example measured magnetometer signatures that were removed because it was confusing as to what precisely was the signature (highly anomalous signatures). This class corresponded to 7% of the removed cases.

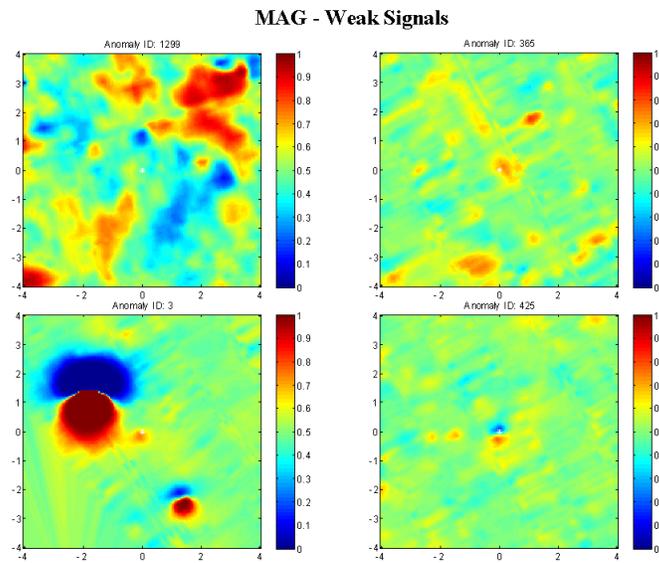


Fig. 42. Example measured magnetometer signatures that were removed because of weak signals in the location of the specified target location (white point). Note that there are sometimes strong nearby targets and it is not clear if there is actually target-location error. This class corresponded to 54% of the removed cases.

IX. Performance Assessment and Cost Assessment

A. Performance criteria

For each anomaly considered, we have provided the output of each of the algorithms considered, where larger outputs are more indicative of UXO, and smaller outputs are more indicative of non-UXO, or clutter, items. The performance criteria are outlined in Figure 43.

Performance Criterion	Description	Primary or Secondary
Analysis Time	Time required to make a decision for each anomaly	Secondary
Fractional probability of detection	Number of detections divided by the number of ordnance in the test site detected by the contractor	Primary
Fractional probability of false alarm	Number of false positives (<i>i.e.</i> , declaration of ordnance) corresponding to clutter items divided by the number of opportunities for false positives (<i>i.e.</i> , clutter or blanks declared by the contractor)	Primary

Fig. 43. Performance criteria for Sibert test.

B. Performance confirmation methods

Analysis time was logged manually (total time per anomaly was approximately 20 minutes, on average). ROC results using Fractional Probability of Detection and Fractional Probability of False Alarm (scored by IDA) are provided below, and relative performance between active and non-active classification techniques is examined. The performance of the various sensor combinations is compared.

C. Data analysis, interpretation and evaluation

Inversion of the data and application of each of the algorithms has been performed. Performance associated with factors associated with each of the classification techniques (*e.g.*, probability density functions utilized for the active-learning techniques, in a supervised and semi-supervised setting) have been examined. All evaluations are described in detail below.

D. Cost reporting

As noted above, the time required to formulate the algorithm output associated with each anomaly considered has been tracked. All of this data have been provided in this report.

E. Cost analysis

The baseline alternative technology to be used as a reference for this site will be established in cooperation with the ESTCP Program office. The anticipated cost basis for this technology demonstration is the number of anomalies that can be classified per unit time, while taking into account classifier training time. The main cost driver of this technology is the time it takes to perform the dipole model inversion. As a result, the cost will be a function of the number of anomalies, and hence also a function of the site size.

X. Classification Results - Non-Active Learning

SIG performed the classification study in two ways: (i) like all performers we designed classifiers based on the labeled data as give to us (excluding, as discussed above, those items that didn't fit the dipole model well); and (ii) we performed active learning to acquire labeled data, assuming no *a priori* labeled data. In this section we present the results of (i), and in Section 11 we present results for (ii). All receiver operating characteristic (ROC) curves presented in this section were computed by the Institute for Defense Analyses (IDA), and the summary tables were computed by IDA and by SIG (the latter after re-analysis of the results after "truth" had been revealed).

A. Presentation format for all IDA-generated ROCs

Concerning the format of the ROC plots below, as computed by IDA, the following format is employed. For each point on the ROC curve, the scoring software calculates and draws 95% vertical confidence intervals around the Pd value using the exact binomial distribution with no adjustment for multiple comparisons. This means that IF one were to ignore the impact of multiple comparisons, THEN one could say with 95% certainty that the Pd value of a particular point on the ROC curve lies within the vertical confidence interval passing through that point. One canNOT form an upper (lower) confidence band by connecting the upper (lower) ends of all points confidence intervals and then say with 95% certainty that the entire ROC curve lies between the upper (lower) confidence bands.

The scoring software colors in each point of the ROC curve (each small dot) based on where the corresponding value of the first cut-off point on item rank lies with respect to the demonstrator-suggested categories:

- Green: point lies within demo-suggested Category 1 (*i.e.*, "likely to be not UXO")
- Yellow: point lies within demo-suggested Category 2 (*i.e.*, "cant decide")
- Orange: point lies within demo-suggested Category 4 (*i.e.*, "cant decide")
- Red: point lies within demo-suggested Category 5 (*i.e.*, "likely to be UXO")

The scoring software plots three large dots on the ROC curve, each specifying one particular value of the first cut-off point on item rank:

- Blue: the value suggested by demonstrators
- Cyan: the value that would have resulted in the minimum FP when $P_d=1.00$
- Magenta: the value that would have resulted in the minimum FP when $P_d=0.95$

B. EM61 sensor

For the EM61 sensor, we provide a detailed discussion of the format of the result presentation, and the results from the other sensors are presented in the same format. In Figures 44 and 45 we present supervised and semi-supervised classifier ROCs, broken down by the SE and SW portions of the Sibert site. Note that in these plots the ROC hugs the $P_d = 0$ for an extended number of false alarms, before starting to quickly rise. The items for which no classification was performed would in practice be excavated, and the initial $P_d = 0$ portion of the ROC corresponds to these excluded items. The P_{fa} at which the ROC starts to rise corresponds to the fraction of non-UXO items for which no classification decision was made. Therefore, one way to interpret these results is to consider performance after the ROC starts to rise (for those P_{fa} for which $P_d > 0$). From this standpoint, we note that the rise in the ROCs is particularly steep for the supervised classifier in Figure 45, indicating that for those items for which classification was performed very encouraging results were achieved. Recall from the discussion in Section 7.1 that the semi-supervised classifier tends to be more conservative than the supervised classifier – this results in more false alarms for a given detection rate. One therefore notes that the semi-supervised results in Figure 44 are good, apart from the excluded items, but these ROCs are not as steep as those associated with the supervised classifier in Figure 45.

The ROC curves in Figures 44 and 45 and below are interesting, but from a practical standpoint one does not get the opportunity to draw ROC curves in the field (because some items are unexcavated, and therefore “truth” is not known for these). Therefore, of more practical importance is classification performance based on the setting of thresholds (the ROC shows results for *all* thresholds, and selection of the threshold in the field is a very important practical issue). In Section 7 we discussed setting thresholds based on a probabilistic risk analysis, and for this

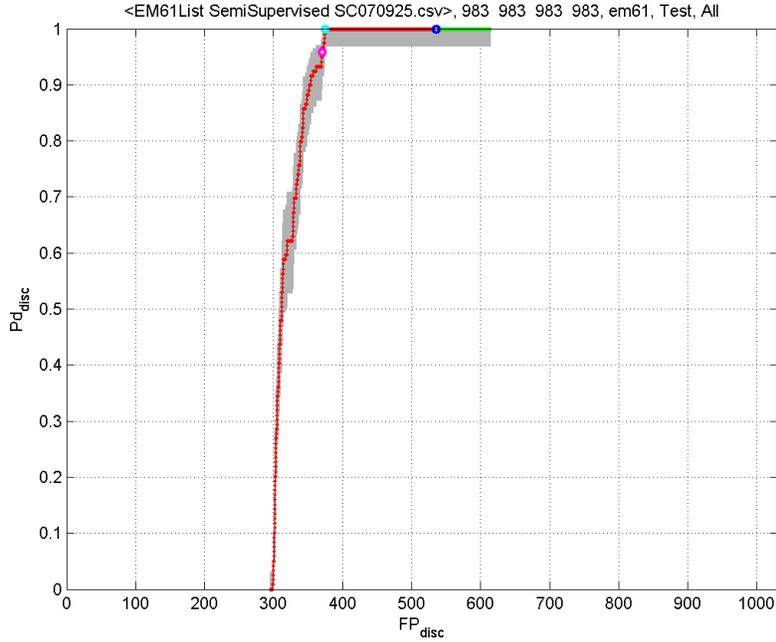


Fig. 44. Receiver operating characteristics (ROCs) for the Sibert site, based on the EM61 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

study SIG submitted three thresholds $C = 100$, $C = 50$, and $C = 25$ (the larger C , the more conservative the threshold setting). In Figures 44 and 45 we present summary charts for the semi-supervised and supervised classifiers, respectively. In these tables, perhaps the most important information is the probability of correct discrimination P_{disc} and the probability of false alarm Pfa_{disc} at the specified threshold T_{demo} (recall that SIG provided three distinct T_{demo}). For the more-conservative semi-supervised classifier, SIG had no missed UXO for any of the thresholds, while for the supervised classifier SIG missed one UXO at all thresholds.

In the tables in Figures 46 and 47 two types of results are presented. On the left are results as scored by IDA, in which they declared all items for which a classification decision was not rendered a false alarm. As indicated in Section 8, SIG was highly (maybe overly) conservative in removing items from classification (all items for which reliable dipole features could not be reliably rendered were not classified), despite the fact that we had significant information to suggest that all of these items were not UXO. Therefore, the right portion of Figures 46

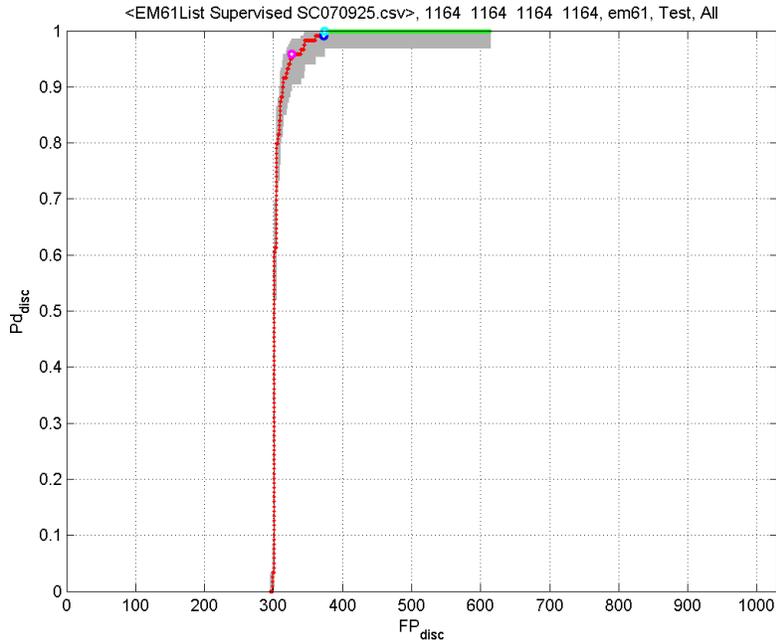


Fig. 45. Receiver operating characteristics (ROCs) for the Sibert site, based on the EM61 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

and 47 present P_{disc} and Pfa_{disc} results only for those items that SIG actually classified via its classifier (only those for which reliable dipole features were extracted). One notes that the Pfa_{disc} in this case is quite low for the (less conservative) supervised classifier: on the SE portion the Pfa_{disc} is 0.17 at the most-conservative threshold setting, and for the SW site it is 0.29 . This demonstrates that the classifier performed quite well on the items for which classification was actually performed (the steep ROCs in Figures 44 and 45 also reflect this).

Summarizing the results in Figures 44- 47, excluding the items for which classification was not attempted, the supervised and semi-supervised classifiers performed well using the EM61 data. As expected, the semi-supervised classifier was more conservative than its supervised counterpart; the semi-supervised classifier missed no UXO at the specified thresholds, albeit with a larger false-alarm rate.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed: 295)						
Alg	Area	C	P_{disc} when $T=T_{demo}$ [FN]	$P_{fa_{disc}}$ when $T=T_{demo}$ [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{demo}$ [FN]	$P_{fa_{disc}}$ when $T=T_{demo}$ [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (semi-sup)	All	100	1.00 [0]	0.95 [583]	0.61 [5, 373]	0.79 [0, 488]	SIG (semi-sup)	All	100	1.00 [0]	0.90 [288,31]	0.24 [5, 78]	0.61 [0, 193]
		50	1.00 [0]	0.90 [553]					50	1.00 [0]	0.81 [258,61]		
		25	1.00 [0]	0.87 [536]					25	1.00 [0]	0.76 [241,78]		

Fig. 46. Summary Sibert performance for the EM61 sensor, using a semi-supervised classifier.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 295)						
Alg	Area	C	P_{disc} when $T=T_{demo}$ [FN]	$P_{fa_{disc}}$ when $T=T_{demo}$ [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{demo}$ [FN]	$P_{fa_{disc}}$ when $T=T_{demo}$ [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (sup)	ALL	100	1.00 [0]	0.60 [372]	0.56 [5, 346]	0.68 [0, 419]	SIG (sup)	All	100	0.99 [1]	0.24 [77,242]	0.16 [5, 51]	0.40 [0, 124]
		50	1.00 [0]	0.60 [366]					50	0.99 [1]	0.22 [71,248]		
		25	1.00 [0]	0.59 [362]					25	0.99 [1]	0.21 [67,252]		

Fig. 47. Summary Sibert performance for the EM61 sensor, using a supervised classifier.

C. Magnetometer

The results for the magnetometer data are similar to those for the EM61, and in some respects slightly better. For example, for the magnetometer data the supervised and semi-supervised classifiers both missed no UXO, at all submitted thresholds. Considering the supervised-classifier

results in Figure 51, at the threshold $C = 100$ (the most conservative setting), all UXO were detected; further, for the SE region 82% of the non-UXO were left unexcavated, with this 62% for the SW site (for the items for which classification was attempted).

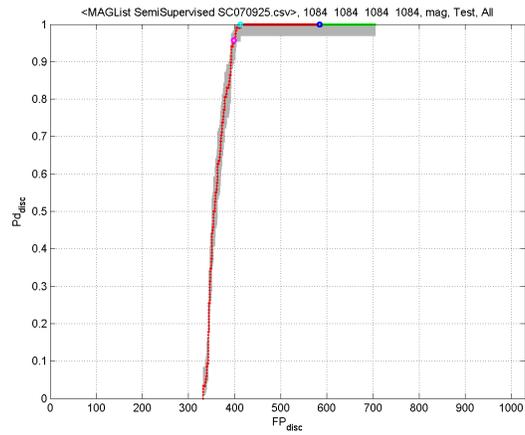


Fig. 48. Receiver operating characteristics (ROCs) for the Sibert site, based on the magnetometer sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

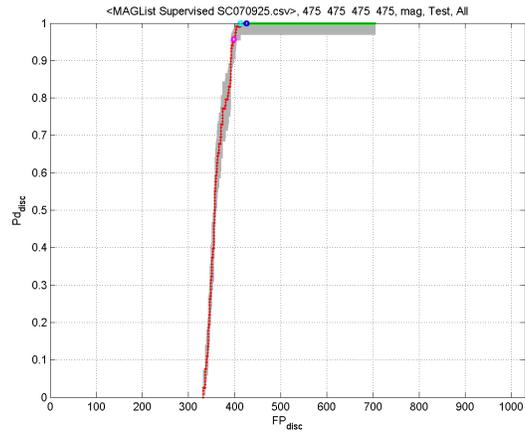


Fig. 49. Receiver operating characteristics (ROCs) for the Sibert site, based on the magnetometer sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 333)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (sup)	All	100	1.00 [0]	0.85 [672]	0.56 [5, 396]	0.58 [0, 406]	SIG (sup)	All	100	1.00 [0]	0.81 [338,233]	0.17 [5, 63]	0.20 [0, 73]
		50	1.00 [0]	0.88 [628]					50	1.00 [0]	0.79 [285,77]		
		25	1.00 [0]	0.83 [583]					25	1.00 [0]	0.67 [250,122]		

Fig. 50. Summary Sibert performance for the magnetometer sensor, using a semi-supervised classifier.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 333)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (sup)	ALL	100	1.00 [0]	0.61 [430]	0.57 [3, 402]	0.58 [0, 406]	SIG (sup)	ALL	100	1.00 [0]	0.26 [97,275]	0.19 [5, 69]	0.20 [0, 73]
		50	1.00 [0]	0.60 [425]					50	1.00 [0]	0.25 [92,280]		
		25	1.00 [0]	0.59 [417]					25	1.00 [0]	0.23 [84,288]		

Fig. 51. Summary Sibert performance for the magnetometer sensor, using a supervised classifier.

D. EM63 sensor

In some respects the EM63 results were the best, although the number of items considered was only roughly 20% as much as considered for the EM61 and magnetometer sensors (see Figure 32). Consistent with the results for the EM61 and magnetometer, the supervised and semi-supervised ROCs are similar, although the semi-supervised ROCs are more conservative. Considering the SE portion, both the supervised and semi-supervised classifier (on the items classified) detected

roughly 80% of the UXO before a single false alarm. On the SW site the semi-supervised classifier actually performed slightly better than the supervised classifier (it is believed that, in this case, the contextual information afforded to the semi-supervised classifier by the unlabeled data improved the classifier decision boundary).

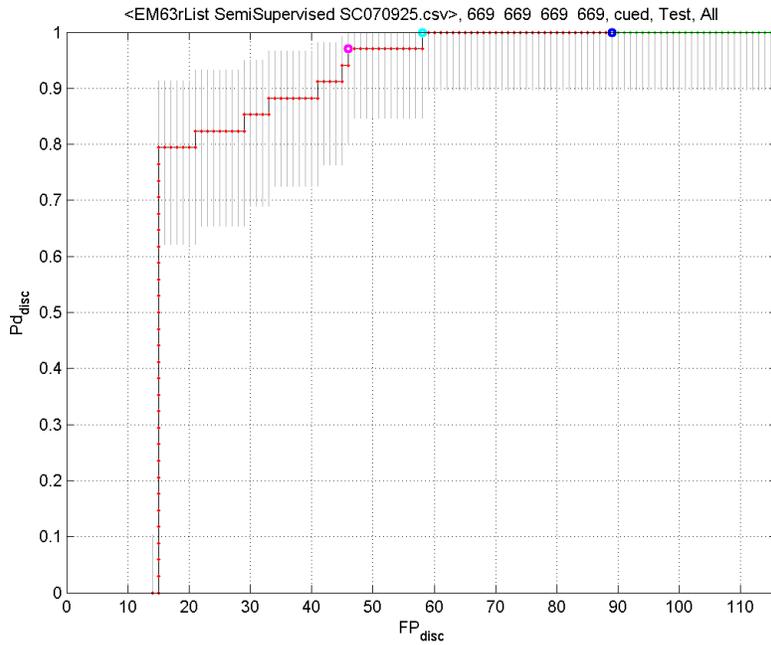


Fig. 52. Receiver operating characteristics (ROCs) for the Sibert site, based on the EM63 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

Considering the summary charts in Figures 54 and 55, no UXO were missed at any of the thresholds. However, the thresholds appeared to be very conservative, in the sense that the ROCs in Figures 52 and 53 are very good, but the number of false alarms at the specified thresholds (Figures 54 and 55) is relatively high. This suggests that, in future studies, the high quality of the EM63 data and associated features should be trusted more, and that less-conservative threshold settings may be considered.

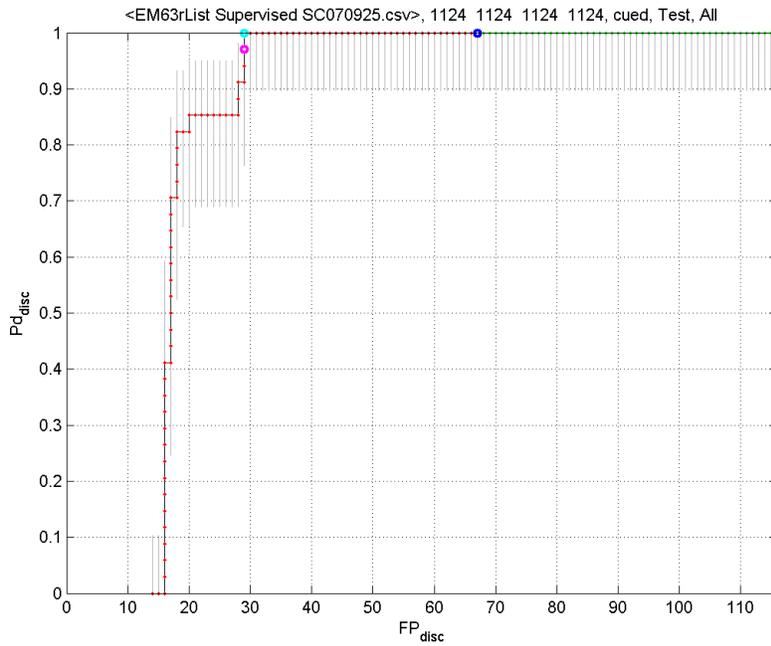


Fig. 53. Receiver operating characteristics (ROCs) for the Sibert site, based on the EM63 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 14)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP, TN]	min $P_{fa_{disc}}$ when $Pd_{disc}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=1.00$ [FN, FP]
SIG (semi-sup)	All	100	1.00 [0]	0.93 [108]	N/A*	0.44 [0, 51]	SIG (semi-sup)	All	100	1.00 [0]	0.92 [94,8]	N/A*	0.36 [0, 37]
		50	1.00 [0]	0.91 [105]					50	1.00 [0]	0.89 [91,11]		
		25	1.00 [0]	0.77 [88]					25	1.00 [0]	0.74 [75,27]		

Fig. 54. Summary Sibert performance for the EM63 sensor, using a semi-supervised classifier.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 14)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (semi-sup)	All	100	1.00 [0]	0.64 [74]	N/A*	0.21 [0, 29]	SIG (semi-sup)	All	100	1.00 [0]	0.58 [60,42]	N/A*	0.15 [0, 15]
		50	1.00 [0]	0.58 [67]					50	1.00 [0]	0.55 [56,46]		
		25	1.00 [0]	0.52 [60]					25	1.00 [0]	0.45 [46,56]		

Fig. 55. Summary Sibert performance for the EM63 sensor, using a supervised classifier.

E. Concatenation of EM63 and magnetometer features

Similar results based on the concatenated features of EM63 and magnetometer are listed below.

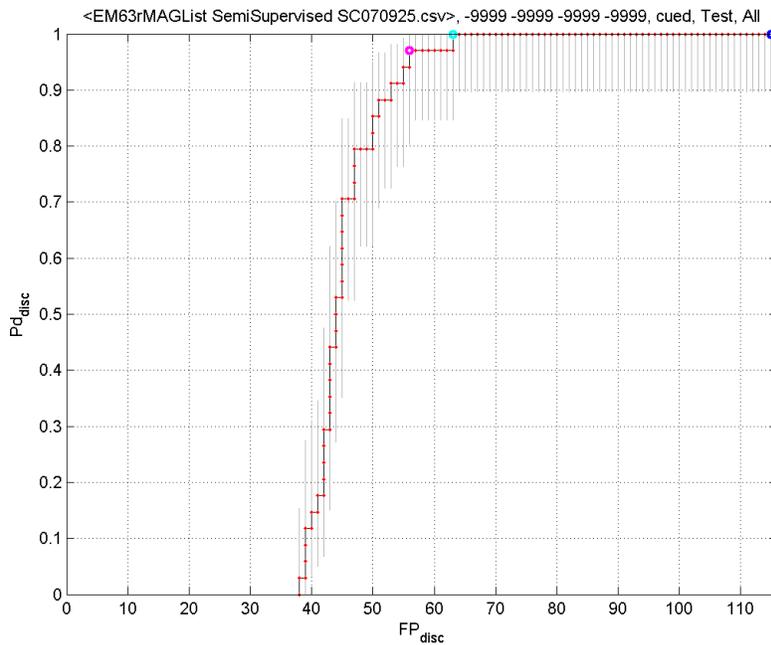


Fig. 56. Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM63 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

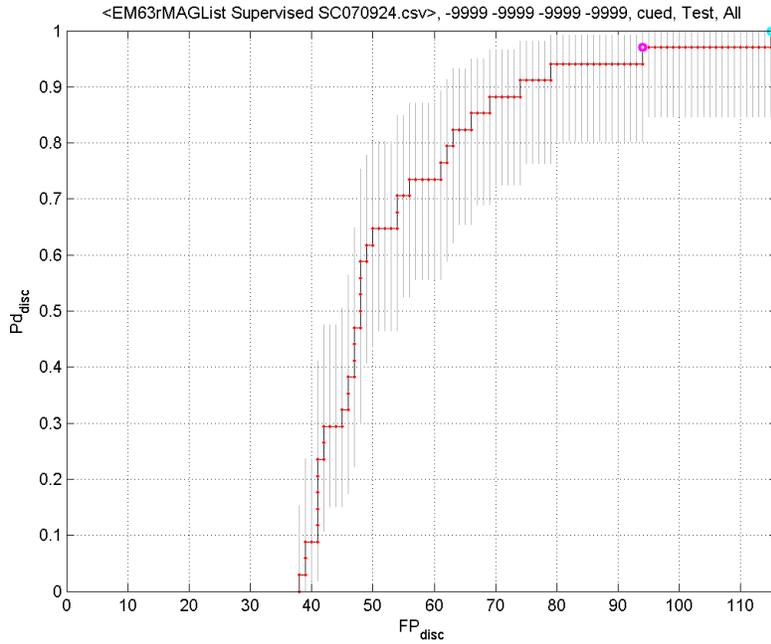


Fig. 57. Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM63 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 2)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ mo [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ mo [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (semi-sup)	All	100	1.00 [0]	1.00 [116]	N/A*	0.38 [0, 56]	SIG (semi-sup)	All	100	1.00 [0]	1.00 [114,0]	N/A*	0.56 [0, 64]
		50	1.00 [0]	0.96 [113]					1.00 [0]	0.97 [111,2]			
		25	1.00 [0]	0.56 [69]					1.00 [0]	0.59 [67,71]			

Fig. 58. Summary Sibert performance for the concatenation EM63 and magnetometer features, using a semi-supervised classifier.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 2)						
Alg	Area	C	P_{disc} when $T=T_{d_{emo}}$ [FN]	$P_{fa_{disc}}$ when $T=T_{d_{emo}}$ [FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{d_{emo}}$ [FN]	$P_{fa_{disc}}$ when $T=T_{d_{emo}}$ [FP, TN]	min $P_{fa_{disc}}$ when $Pd_{disc}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $Pd_{disc}=1.00$ [FN, FP]
SIG (sup)	All	100	0.65 [12]	0.44 [5]	N/A**	0.95 [0, 110]	SIG (sup)	All	100	0.65 [12]	0.42 [49,65]	N/A**	0.84 [0, 98]
		50	0.58 [14]	0.41 [48]					50	0.44 [19]	0.40 [46,68]		
		25	N/A*	N/A*					25	N/A*	N/A*		

Fig. 59. Summary Sibert performance for the concatenation EM63 and magnetometer features, using a supervised classifier.

F. Concatenation of EM61 and magnetometer features

One notes from Figures 60 and 61 that the concatenation of the features from the EM61 and magnetometer yielded very steeply increasing ROC curves, for the items that were actually classified; this was true for both the supervised and semi-supervised classifiers. However, we note that a large fraction of the items were not classified. This was because the results in Figures 60 and 61 only correspond to those items for which both EM61 and magnetometer data were collected (this is a subset of either data set), and of this intersection unfortunately a relatively large fraction were in the set for which classification was not attempted. Nevertheless, when classification was performed, very good performance was manifested.

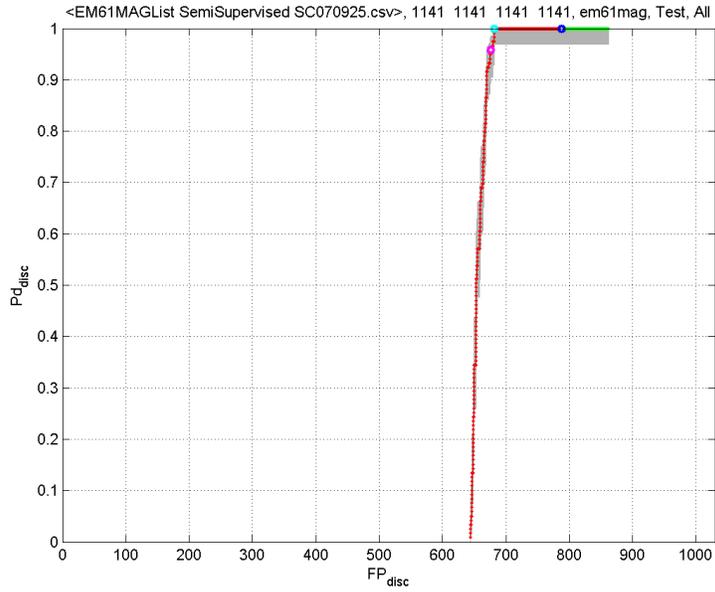


Fig. 60. Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM61 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

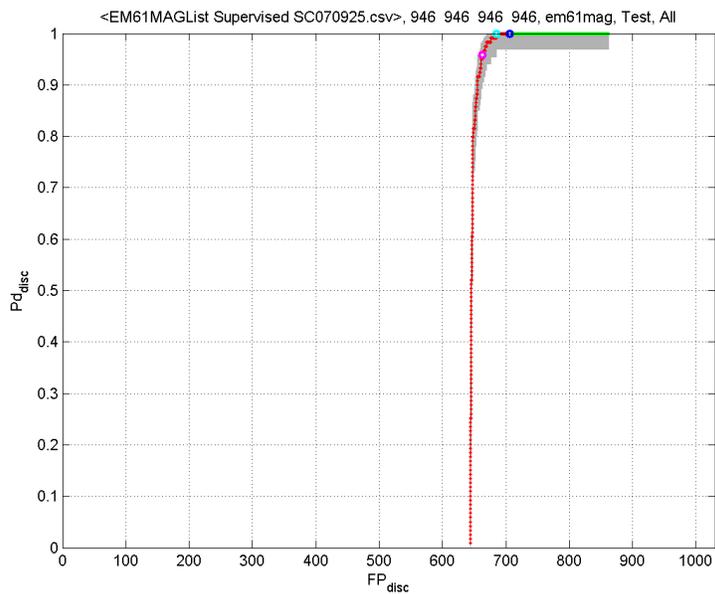


Fig. 61. Receiver operating characteristics (ROCs) for the Sibert site, based on the concatenation of EM61 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 108)						
Alg	Area	C	P_{disc} when $T=T_{danno}$ [FN]	$P_{fa, disc}$ when $T=T_{danno}$ [FP]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{db}$ mo [FN]	$P_{fa, disc}$ when $T=T_{danno}$ [FP, TN]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]
SIG (semi-sup)	All	100	1.00 [0]	0.94 [811]	0.78 [5, 674]	0.79 [0, 680]	SIG (semi-sup)	All	100	1.00 [0]	0.93 [703, 51]	0.75 [5, 566]	0.76 [0, 572]
		50	1.00 [0]	0.91 [787]					50	1.00 [0]	0.90 [679, 75]		
		25	1.00 [0]	0.88 [760]					25	1.00 [0]	0.86 [662, 102]		

Fig. 62. Summary Sibert performance for the concatenation EM61 and magnetometer features, using a semi-supervised classifier.

Original Results Reported by IDA DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 108)						
Alg	Area	C	P_{disc} when $T=T_{db}$ mo [FN]	$P_{fa, disc}$ when $T=T_{danno}$ [FP]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{db}$ mo [FN]	$P_{fa, disc}$ when $T=T_{danno}$ [FP, TN]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]
SIG (sup)	All	100	1.00 [0]	0.82 [705]	0.77 [5, 664]	0.79 [0, 681]	SIG (sup)	All	100	1.00 [0]	0.79 [587, 157]	0.74 [5, 556]	0.76 [0, 573]
		50	1.00 [0]	0.81 [701]					50	1.00 [0]	0.79 [583, 161]		
		25	1.00 [0]	0.81 [695]					25	1.00 [0]	0.78 [586, 168]		

Fig. 63. Summary Sibert performance for the concatenation EM61 and magnetometer features, using a supervised classifier.

G. GEM3 sensor

Note from Figure 32 that none of the GEM3 data were excluded. Therefore, the data quality as given was good for all items considered by the GEM3 sensor. However, as discussed above, it is felt that the spatial sampling was too coarse to render good model fits. Therefore, as anticipated, the classification performance for the GEM3 is far inferior to that of the EM61, magnetometer and EM63. It should be emphasized that this is unlikely to be a consequence of the sensor itself,

but rather because the data were collected at too coarse a spatial sample rate; better performance is anticipated if finer spatial sampling was considered.

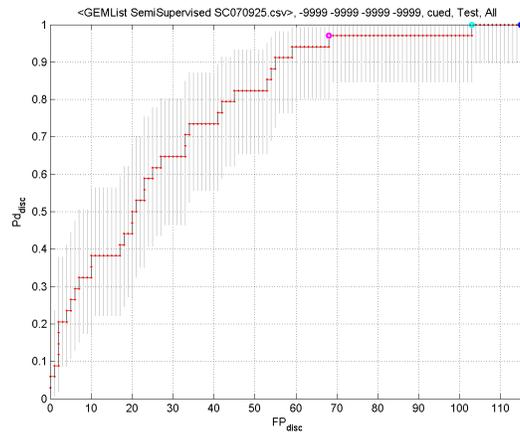


Fig. 64. Receiver operating characteristics (ROCs) for the Sibert site, based on the GEM3 sensor. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

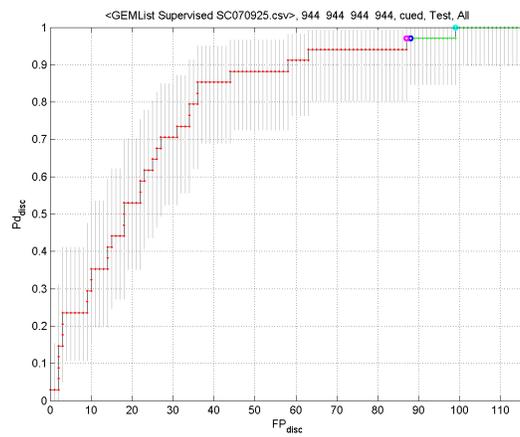


Fig. 65. Receiver operating characteristics (ROCs) for the Sibert site, based on the GEM3 sensor. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

Original Results Reported by IDA DISCRIMINATION							Original Results Reported by IDA DISCRIMINATION						
Alg	Area	C	P_{disc} when $T=T_{db}$ no [FN]	$P_{fa, disc}$ when $T=T_{db, no}$ [FP]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{db, no}$ [FN]	$P_{fa, disc}$ when $T=T_{db, no}$ [FP]	min $P_{fa, disc}$ when $P_{d, disc}=0.95$ [FN, FP]	min $P_{fa, disc}$ when $P_{d, disc}=1.00$ [FN, FP]
SIG (semi- sup)	All	100	1.00 [0]	1.00 [116]	N/A**	0.91 [0, 106]	SIG (sup)	All	100	0.95 [1]	0.82 [85]	N/A*	0.77 [0, 89]
		50	N/A*	N/A*					50	0.95 [1]	0.76 [88]		
		25	1.00 [0]	1.00 [115]					25	0.90 [2]	0.70 [81]		

Fig. 66. Summary Sibert performance for the GEM3 sensor.

H. Concatenation of GEM3 and magnetometer features

The number of items for which both GEM3 and magnetometer data were available was relatively small. Moreover, as discussed above, the quality of the features extracted for the GEM3 was relatively poor. We present results here for concatenated features from the EM63 and magnetometer, to meet the requirements of the DemVal, but for the reasons articulated these results are of limited value.

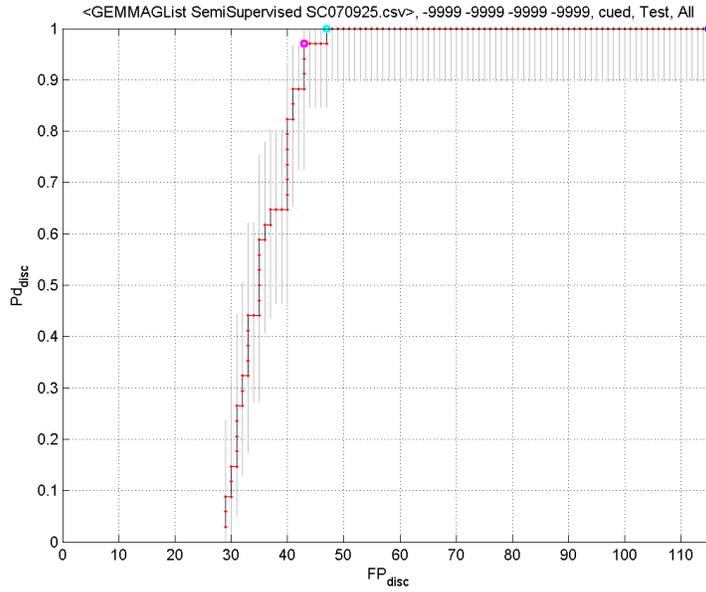


Fig. 67. Receiver operating characteristics (ROCs) for the Sibert site, based on concatenation of the GEM3 and magnetometer features. The results are for a semi-supervised classifier. The form of the plots is discussed in Section 10.1.

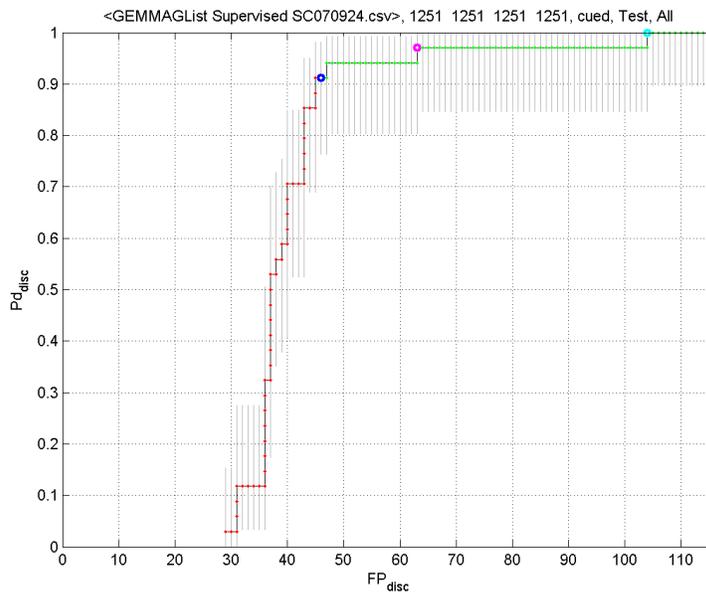


Fig. 68. Receiver operating characteristics (ROCs) for the Sibert site, based on concatenation of the GEM3 and magnetometer features. The results are for a supervised classifier. The form of the plots is discussed in Section 10.1.

XI. Active-Learning Classification Results

Active learning was performed by SIG in cooperation with Dr. Herbert Nelson (of NRL). The test was performed as follows. SIG completely sequestered the label information from the SIG person who performed the active learning, and therefore the test was performed as if we initially had no labels at all. Using the information-theoretic techniques discussed in [8], we determined roughly ten items for which label information would be most informative. This list was sent via email to Dr. Nelson, who then emailed the labels for these items. Using these labels the classifier was trained, and we then asked which additional labels would be most informative for classifier design, with that list again sent to Dr. Nelson. This process was iterated until the information-theoretic measure indicated that there was no further information to be accrued by acquisition of additional labels. In consultation with Dr. Anne Andrews of ESTCP, it was agreed that the active learning should be applied to the data for which both EM61 and magnetometer sensor data are available, since this may be the most common sensor combination in current practice. However, in retrospect this may have been an unfortunate choice, since a large fraction of these data were excluded from the classification study (see the discussion in Section 10.6). In total, labels were requested for 58 items, and of these 14 corresponded to UXO. By considering Figure 32, we note that the number of labeled items used in the active-learning designed classifier is significantly smaller than that used for the traditional learning (in which ESTCP provided labeled data to the performers).

A. Intersection of EM61 and magnetometer data

The performance of the active-learning algorithm on the intersection of EM61 and magnetometer data is as good as that for the non-active-learning results in Figures 60 and 61, despite the reduced number of labeled training samples.

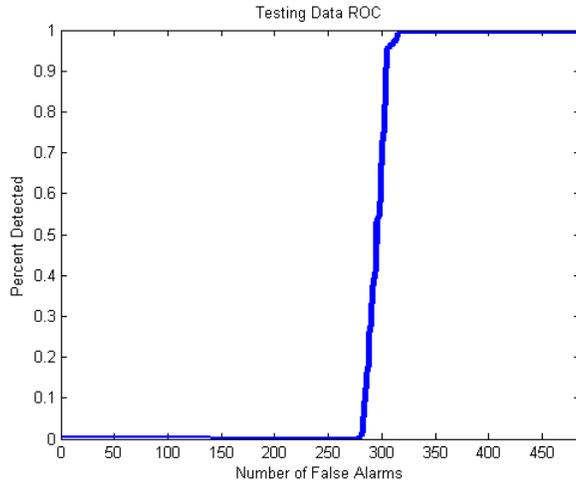


Fig. 69. Receiver operating characteristics (ROCs) for the Sibert site, based on concatenated features from the EM61 and magnetometer sensors. The results are for a semi-supervised classifier, and are based on labeled data acquired via active learning.

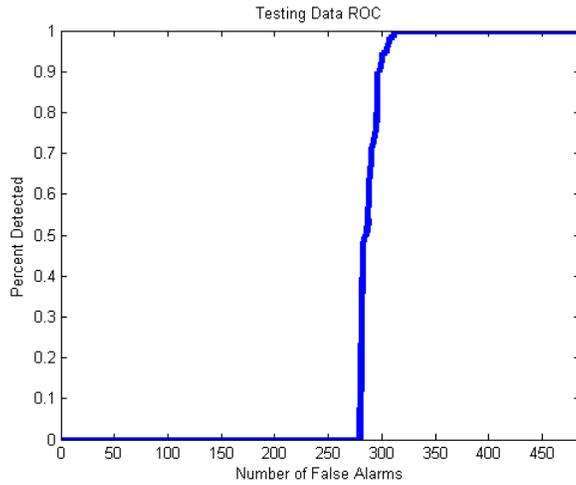


Fig. 70. Receiver operating characteristics (ROCs) for the Sibert site, based on concatenated features from the EM61 and magnetometer sensors. The results are for a supervised classifier, and are based on labeled data acquired via active learning.

Original Results DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 279)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ no [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ no [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ o [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (semi-sup, active)	All	100	1.00 [0]	0.87 [421]	0.59 [6, 305]	0.61 [0, 317]	SIG (semi-sup, active)	All	100	1.00 [0]	0.67 [142,64]	0.13 [6, 26]	0.18 [0, 38]
		50	1.00 [0]	0.80 [388]					50	1.00 [0]	0.52 [107,99]		
		25	1.00 [0]	0.74 [361]					25	1.00 [0]	0.40 [82,124]		

Fig. 71. Summary Sibert performance for concatenated features from the EM61 and magnetometer sensors, based on a semi-supervised classifier. The labeled data were acquired via active learning.

Original Results DISCRIMINATION							Results Considering Objects Where Discrimination Attempted DISCRIMINATION (Excluded Anomalies Removed 279)						
Alg	Area	C	P_{disc} when $T=T_{disc}$ [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ [FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]	Alg	Area	C	P_{disc} when $T=T_{disc}$ o [FN]	$P_{fa_{disc}}$ when $T=T_{disc}$ [FP, TN]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=0.95$ [FN, FP]	min $P_{fa_{disc}}$ when $P_{d_{disc}}=1.00$ [FN, FP]
SIG (sup, active)	All	100	1.00 [0]	0.60 [311]	0.12 [6, 300]	0.60 [0, 311]	SIG (sup, active)	All	100	1.00 [0]	0.15 [32,178]	0.10 [6, 21]	0.15 [0,32]
		50	1.00 [0]	0.60 [309]					50	1.00 [0]	0.15 [30,176]		
		25	0.99 [1]	0.59 [307]					25	0.99 [1]	0.14 [28,174]		

Fig. 72. Summary Sibert performance for concatenated features from the EM61 and magnetometer sensors, based on a supervised classifier. The labeled data were acquired via active learning.

B. Individual EM61 and magnetometer processing (post analysis)

In an analysis by SIG, after “truth” was revealed, the labeled data acquired for the intersection of the EM61 and magnetometer data were applied to design classifiers separately for the EM61

and magnetometer (in some sense this is “unfair” to the individual EM61 and magnetometer classifiers, since the actively acquired labeled data were acquired for the intersection of the sensor data, not for either alone). This test was performed because the fraction of excluded items for the EM61 and magnetometers alone is smaller than that for the intersection of the two combined. Results are presented in Figure 73; apart from the (smaller fraction of) excluded items, the active-learning ROC performance is quite good on the EM61 and magnetometer data alone – very steep ROCs.

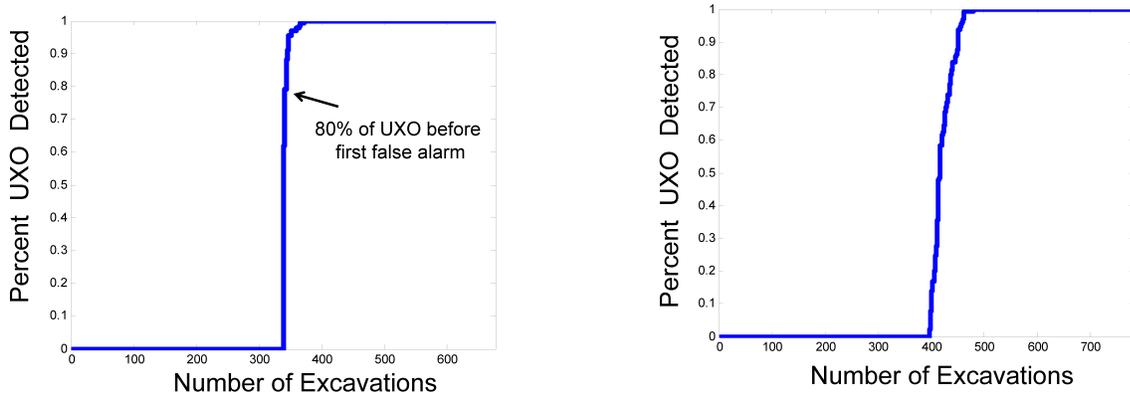


Fig. 73. Receiver operating characteristics (ROCs) for the EM61 (left) and magnetometer (right) sensors alone. The results are for a supervised classifier (the ROCs for a semi-supervised classifier are virtually identical), and are based on labeled data acquired via active learning.

XII. Cost Assessment

A. Cost breakdown

The costs associated with the analysis of the Sibert data can be broken down into the following categories: data preparation, anomaly definition, feature extraction, classifier training/testing, active learning, and results reporting. While no algorithm development was necessary under this effort, time was required to adapt existing algorithms to the current data set and then quality assurance measures were taken to ensure that the adapted algorithms were appropriately configured. This cost section generally defines the costs of each of the necessary steps to process the various data sets from the Sibert study. The costs below do not include any costs related to planning meetings, reports generation, or travel.

Data Preparation: Data preparation includes time required to configure code to read in and organize the data set, segregate training and testing data and adapt existing software. This is a data set dependent process and is independent of the number of the amount of data or the number of anomalies.

Anomaly Definition: Designation of data points associated with each anomaly. Determining the data belonging to each anomaly is usually defined using a manual or automated technique to define the anomaly boundary. If manually defined, each anomaly must be visually inspected and a software tool is used to designate the polygonal boundary. This process is very time consuming, but ensures the highest accuracy of the data points belonging to each anomaly. The automated approach is usually faster, but may assign some data points incorrectly (for example when two anomalies are very close to each other). The anomaly boundaries in the current study were selected automatically and then reviewed and validated by a person. The cost for this process is linearly dependent on the number of anomalies.

Feature Extraction: Feature extraction involves the computationally intensive application of an appropriate parametric data model and extraction of parameters that capture the inherent physical or electrical characteristics of each anomaly. This process can be time consuming due to the large number of parameters typically in the data model as well as their non-linear nature. The

non-linear nature of the models necessitates the random initialization of many solution attempts to find the overall best local solution. The time required is a linear function of the number of anomalies present in the data set.

Classifier Training and Testing: Once features are extracted, visualization of the data features occurs to ensure that the data are reasonably distributed. Using any available training data, classifier parameters are appropriately set (whether supervised or semi-supervised) by performing Leave-One-Out (LOO) cross-validation on the training data. In general the cost of this process is not dependent on the number of anomalies

Active Learning: Active learning is an iterative process whereby an information metric is computed for each unlabeled data point to identify the next most informative labels to acquire. The process can be time consuming if the procedure for acquiring labels is delayed significantly at each label request. The number of label request iterations is the driving cost factor for this process and is dependent on the statistics of the data set.

Results Reporting: Once the results are computed, Receiver Operating Characteristic (ROC) curves and other performance metrics (risk analysis, misses, etc) are computed and graphs and tables generated to highlight the results. This process is not dependent on the number of anomalies in the data set.

B. Cost Benefit

The cost benefit for discrimination is based on the number of anomalies that do not need to be excavated or at least can be excavated in different ways depending on the probability that the anomaly represents a UXO or whether it is more likely to be clutter. In terms of the cost benefit of the supervised classification approach or the semi-supervised classification approach, they both cost the same to apply to the data set; therefore there is no cost differentiator. However, the assumptions and the risk trade-off in determining the UXO probability of an anomaly vary between the two approaches and impact the likelihood of the supervised approach leaving more objects in the ground versus the semi-supervised approach but potentially increasing the risk of missing a UXO.

	GEM3	GEM+MAG	EM61	EM61+MAG	EM63	EM63+MAG	MAG	TOTAL
Data Preparation	\$2		\$2		\$2		\$2	\$8
Anomaly Definition	\$4		\$4		\$4		\$4	\$14
Feature Extraction	\$8		\$8		\$8		\$8	\$32
Supervised Classifier Training/Testing	\$3	\$3	\$3	\$3	\$3	\$3	\$3	\$21
Semisup Classifier Training/Testing	\$3	\$3	\$3	\$3	\$3	\$3	\$3	\$21
Active Learning			\$2	\$2			\$2	\$6
Results Reporting	\$2	\$2	\$2	\$2	\$2	\$2	\$2	\$11
[values are in thousands]	\$21	\$8	\$23	\$10	\$21	\$8	\$23	\$113

Fig. 74. Cost model for work performed during the Camp Sibert discrimination study. Active learning was only performed on a subset of the sensor data provided.

The most important element of the discrimination approach that drives cost is the decision to employ active learning or not. While active learning costs more in terms of the time it takes to iterate on label selection, the overall number of training anomalies required to train the classifier can be significantly lower with active learning. To illustrate, the number of labels provided by the program office to train the EM61 was 127. The 127 training anomalies were selected based on emplaced UXO and a sampling of the anomalies in the Sibert range. Using this method to provide training data, there is no measure for the completeness of the representation of the training data in the feature space. Active learning for the EM61, however, required only 58 training data points and provides an informative measure of the benefit of adding any additional data points. Active learning directs which new labels to acquire based on the statistics of the data in the feature space. Active learning can provide significant cost savings by allowing for fewer training data excavations.

REFERENCES

- [1] Y. Zhang, L. Collins, H. Yu, C. Baum, and L. Carin, "Sensing of unexploded ordnance with magnetometer and induction data: Theory and signal processing," *IEEE Transactions of Geoscience and Remote Sensing*, vol. 41, pp. 1005–1015, 2003.
- [2] —, "Sensing of unexploded ordnance with magnetometer and induction data: theory and signal processing," *IEEE Trans. Geoscience & Remote Sensing*, 2003.
- [3] Q. Lu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semi-supervised and active learning," *IEEE Trans. Geoscience & Remote Sensing*, 2008 (to appear).
- [4] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: application to sensing subsurface ux0," *IEEE Trans. Geoscience & Remote Sensing*, 2004.
- [5] L. Carin, H. Yu, Y. Dalichaouch, A. Perry, P. Czipott, and C. Baum, "On the wideband EMI response of a rotationally symmetric permeable and conducting target," *IEEE Transactions of Geoscience and Remote Sensing*, vol. 39, pp. 1206–1213, 2001.
- [6] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.
- [7] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.
- [8] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: application to sensing subsurface ux0," *IEEE Transactions of Geoscience and Remote Sensing*, vol. 42, pp. 2535–2543, 2004.