

FINAL REPORT

Cost-Aware Design of a Discrimination Strategy for Unexploded Ordnance Cleanup

SERDP SEED Project MR-1715

FEBRUARY 2011

Jeremiah Remus
Clarkson University

This document has been cleared for public release



This report was prepared under contract to the Department of Defense Strategic Environmental Research and Development Program (SERDP). The publication of this report does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official policy or position of the Department of Defense. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the Department of Defense.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The 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 the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Services and Communications Directorate (0704-0188). 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 ORGANIZATION.					
1. REPORT DATE (DD-MM-YYYY) 25-02-2011		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 01-12-2009 - 31-3-2011	
4. TITLE AND SUBTITLE Cost-Aware Design of a Discrimination Strategy for Unexploded Ordnance Cleanup				5a. CONTRACT NUMBER W912HQ-10-C-0002	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Dr. Jeremiah Remus Dr. Leslie Collins Dr. Stacy Tantom Dr. Kenneth Morton				5d. PROJECT NUMBER MR-1715	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Clarkson University 8 Clarkson Ave. Potsdam, NY 13699				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) SERDP PROGRAM OFFICE 901 NORTH STUART STREET SUITE 303 ARLINGTON VA 22203				10. SPONSOR/MONITOR'S ACRONYM(S) SERDP	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The objective of this project was to conduct a preliminary investigation into the potential benefit of awareness of the specific performance criterion (100% UXO detection) in each stage of the UXO discrimination processing strategy. This project consisted of several large-scale classification studies to carefully analyze the performance of different classification algorithms and the effects of training data when operating at 100% UXO detection. The various classification algorithms included in this study provide a diverse representation of the different theoretical approaches to pattern classification, and allow for comparison of the effect of different classifier properties on performance at the 100% detection operating point. This study provides evidence that the desire to operate at 100% detection may lead to a preference for certain algorithms in the different stages of the discrimination strategy. Careful consideration and selection of methods used in each stage of discrimination strategy may greatly impact performance at the 100% detection goal.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES 32	19a. NAME OF RESPONSIBLE PERSON Dr. Jeremiah Remus
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code) (315) 268-2126

Table of Contents

Table of Contents	ii
List of Figures	iii
List of Tables	iii
Keywords	iv
List of Acronyms	iv
Acknowledgements	iv
Abstract	1
Objective	2
Background	2
General UXO discrimination strategy framework	4
Relevant theory	6
Materials and Methods	8
Results and Discussion	11
Task #1: Classifier design	11
Task #2: Sensitivity to feature selection and model inversion	17
Conclusions and Implications for Future Research	21
Literature Cited	23
Appendices	26
MATLAB Code: Model inversion outlier identification metric	26
MATLAB Code: Distance likelihood ratio test (DLRT) classifier	27

List of Figures

Figure 1. General framework of UXO/clutter discrimination strategy, from raw data to “dig / no dig” declaration.....	5
Figure 2. Representation of the nine-dimensional feature space in two dimensions using principal components analysis.	9
Figure 3. Boxplot showing distribution of P_{FA} at $P_D = 1$ over 1000 sets of training/test data for six classifiers.	11
Figure 4. Performance rankings for each classifier algorithm over all 1000 training/test sets corresponding to the results presented in Figure 3.	12
Figure 5. Performance rankings for each classifier algorithm over all 1000 training/test sets in the experiment testing classifier propensity to declare UXO.	13
Figure 6. Multidimensional scaling representation of the classifier algorithm outputs for five distinct test sets (corresponding to the five subplots).	14
Figure 7. Boxplot showing distribution of P_{FA} at $P_D = 1$ over 1000 sets of training/test data for six classifiers when the amount of available training data is significantly reduced (100 anomalies, at least 20 of which are UXO).	15
Figure 8. Illustration of the variability in the synthetic data set introduced by increasing the noise σ_n	16
Figure 9. <i>Left</i> : Effects of mismatch in the training and test data, represented by the mean P_{FA} at $P_D = 1$ (y-axis) and standard deviation of P_{FA} at $P_D = 1$ (x-axis). <i>Right</i> : Effect of DLRT parameter K on PF at $P_D = 1$, as measured based on 10-fold cross-validation use the baseline data.....	17
Figure 10. Effects of mismatch between training and test data, expressed using statistics of distribution of P_{FA} at $P_D = 1$, when the training set size is increased to $N = 500$ (left) and $N=1000$ (right).	18
Figure 11. Boxplots showing distribution of P_{FA} at $P_D = 1$ using seven feature sets of increasing dimensionality with four different classifiers.	19
Figure 12. Matrix of correlations between various performance measures calculated within and across two independent test sets.	20
Figure 13. <i>Left</i> : Scatter plot of the outlier identification metric A for each anomaly. <i>Right</i> : Scatter plot of the features in a two-dimensional space (i.e. Figure 2) with potential outliers identified via the outlier identification metric.	21

List of Tables

Table 1. List of classifier algorithms and relevant properties.....	10
---	----

Keywords

Unexploded ordnance, operating-point aware, pattern classification techniques, Monte Carlo experiments

List of Acronyms

ANN: Artificial Neural Network
AUC: Area Under the Curve
BRAC: Base Realignment And Closure
DLRT: Distance Likelihood Ratio Test
EER: Equal Error Rate
EMI: Electromagnetic Induction
FLD: Fisher Linear Discriminant
FUDS: Formerly Used Defense Sites
GLRT: Generalized Likelihood Ratio Test
GMM: Gaussian Mixture Model
KNN: K Nearest Neighbor
MM/RMP: MetalMapper / RobustMultiPrince features
 P_D : Probability of Detection
 P_{FA} : Probability of False Alarm
POMDP: Partially-Observable Markov Decision Process
RF: Random Forest
RMS: Root-Mean-Square
ROC: Receiver-Operating Characteristic
RVM: Relevance Vector Machine
SON: Statement of Need
SVM: Support Vector Machine
TEM: Time-domain Electromagnetic
UXO: Unexploded ordnance

Acknowledgements

This work was completed in collaboration with co-performers at Duke University: Dr. Leslie Collins, Dr. Stacy Tantum, and Dr. Kenneth Morton. Technical support from Skip Snyder at Snyder Geosciences, in the form of data, features, and technical advice, was greatly appreciated. This project was fully funded by SERDP under project MR-1715.

Abstract

Objectives: Cleanup of subsurface unexploded ordnance (UXO) at military installations and training ranges is an expensive and time-consuming challenge. While the goals in UXO remediation are very clear, to cleanup all UXO as efficiently as possible, little effort has focused on designing robust and efficient signal processing strategies with the specific performance goal dictated by the regulators in mind. This project originated from the hypothesis that performance and robustness may be improved over the classical approaches by specifically considering the desired operating point of the UXO discrimination strategy (100% detection) during the construction of each stage of the signal processing sequence that is needed to make the “dig/no dig” decision. From a statistical decision theory perspective, operating at this specific point has implications that may impose a strong preference for certain processing techniques in the UXO/clutter discrimination process. The objective of this project was to conduct a preliminary investigation into the potential benefit of awareness of the specific performance criterion (100% UXO detection) in each stage of the UXO discrimination processing strategy. This work should lead to new strategies for training and classification, and may suggest guidelines for all stages of data processing.

Technical Approach: This project consisted of several large-scale classification studies to carefully analyze the performance of different classification algorithms and the effects of training data when operating at 100% UXO detection. The data used in this study was collected during the Camp San Luis Obispo demonstration with the MetalMapper TEM sensor. The various classification algorithms included in this study provide a diverse representation of the different theoretical approaches to pattern classification, and allow for comparison of the effect of different classifier properties on performance at the 100% detection operating point.

Results: Across a large number of experiments, strong performance was consistently observed with a nonparametric classification algorithm that makes decisions locally in feature space based on neighboring training samples. Such a classifier shares properties with the library-matching classifiers that are often used in the UXO research community for classification based on the polarizability curves. Additionally, preliminary analysis of a method for evaluating the outputs of the model inversion procedure shows potential for identifying potential outliers (which drive performance at the 100% detection operating point) for more careful follow-on analysis.

Benefits: This study provides evidence that the desire to operate at 100% detection may lead to a preference for certain algorithms in the different stages of the discrimination strategy. Careful consideration and selection of methods used in each stage of discrimination strategy may greatly impact performance at the 100% detection goal. This study provides preliminary work towards guidelines for classifier design, use of training data, model inversion, and feature selection in the UXO discrimination algorithm that will eventually lead to more robust methods of data processing.

Objective

The objective of the work conducted in support of MR-1715 was to determine the degree of benefit achieved when using a “performance operating-point aware” approach to UXO discrimination. A series of experiments were conducted to investigate two basic research thrust areas: (1) a re-consideration of the classifier methods and training set design for operating at 100% UXO detection, and (2) development of guidelines for the data pre-processing, model inversion, and feature selection based on quantification of the sensitivity of the 100% UXO detection operating point to these prerequisite data processing stages. This investigation sought to assess the impact of specifically considering the performance criteria required by the regulators (100% UXO detection) during the design of each stage of the UXO discrimination processing strategy. From a statistical decision theory perspective, operating at $P_D=100\%$ has implications that may impose a strong preference for certain processing techniques in the UXO/clutter discrimination process. The goal of this work was to produce a proof-of-concept that adopting this perspective when designing a UXO/clutter discrimination strategy would suggest favoring or avoiding certain methods and techniques in the end-to-end data processing strategy.

This SEED project addresses Statement of Need (SON) MMSEED-10-01. There is a clear need to effectively and cost-efficiently remediate UXO contaminated lands, rendering them safe for their current or intended civilian uses. Understandably, the UXO regulatory cleanup community is strongly averse to the possibility of leaving behind UXO; it is a significant liability when land committed to public use is later found to be contaminated. Therefore, to ensure efficient performance at the desired operating point for the UXO/clutter discrimination strategy (find all of the UXO while at the same time reducing the number of false alarms), it may be necessary to take the specific performance goals into consideration during the design of the end-to-end discrimination strategy. This proposed basic research program was to serve as a proof-of-concept to support our hypothesis that an awareness of the 100% UXO detection operating point during design of the end-to-end discrimination strategy will lead to guidelines and preferences for certain techniques and algorithms at many stages of the overall discrimination strategy.

Background

There are many areas in the United States and throughout the world that are contaminated by or potentially contaminated by unexploded ordnance. In the United States alone there are 1900 Formerly Used Defense Sites (FUDS) and 130 Base Realignment and Closure (BRAC) installations that need to be cleared of UXO. Using current technologies, the cost of identifying and disposing of UXO in the United States is estimated to range up to \$500 billion. Site specific clearance costs vary from \$400/acre for surface UXO to \$1.4 million/acre for subsurface UXO [1]. These approaches, however, usually require significant amounts of human analyst time, and thus those additional costs, which are currently necessary parts of ongoing demonstrations, are

not factored into these numbers. Thus, there is a clear need to effectively and cost-efficiently remediate UXO contaminated lands, rendering them safe for their current or intended civilian uses. Development of new UXO detection technologies with improved data analysis has been identified as a high priority requirement for over a decade.

Several sensor modalities have been explored for the detection and identification of surface and buried UXO. These include electromagnetic induction (EMI), magnetometry, radar, and seismic sensors. These sensors generally experience little difficulty detecting UXO, thus detection does not create the bottleneck that results in the high cost of remediating sites. The primary contributor to the costs and time associated with remediating a UXO-contaminated site is the high false-alarm rate caused by the significant amounts of non-UXO clutter and shrapnel typically found on battlefields and military ranges. A significant contributing factor in the high false alarm rate is the requirement for high confidence in the removal of all UXO during site cleanup.

On sites where anomalies are well separated, statistical signal processing algorithms that exploit recent advances in sensor design and phenomenological modeling have been successfully employed and substantial improvements in performance over traditional “mag and flag” approaches have been demonstrated [2-7]. Recent results from the former Camp San Luis Obispo and Camp Butner demonstrations clearly demonstrate that good discrimination can be effected, and in the case of the Camp Sibert demonstration, all UXO could be identified with a substantial reduction in the number of “dig” declarations. Using nomenclature from decision theory, we can refer to this operating point as the $P_D = 1$ operating point, which corresponds to a UXO probability of detection (P_D) equaling unity. This is the desired operating point in the UXO discrimination scheme; results seen at recent demonstrations indicate that the sensors and algorithms have matured to the point that, under relatively benign test conditions, it can be considered more specifically. The UXO regulator community is highly averse to the possibility of leaving behind UXO; it is a significant liability when land committed to public use is later found to be contaminated. Thus, it is recommended that the research community recognize and adopt this operating point as their standard for measuring performance, such that they are able to meet the needs of the UXO cleanup community. If the technology is incapable of meeting regulators needs, then they will not be willing to adopt it.

Since government regulators require reliable assurances that all UXO have been cleared from a site during cleanup operations, the development of algorithms for discriminating UXO from clutter (reducing cleanup expenses) may benefit if such operating conditions are specifically considered during the design phase. Standard classical approaches to detection and discrimination are not guaranteed to perform well at the $P_D = 1$ operating point. The $P_D = 1$ operating point is defined by the last UXO to be found; thus performance is constrained by the most “anomalous” or outlier UXO in the target set. This perspective could significantly alter the framework of the UXO discrimination strategies, by focusing specifically on robustly identifying the most difficult UXO target. From a statistical decision theory perspective, operating at the P_D

= 1 point has implications that may impose a strong preference for certain processing techniques in the UXO/clutter discrimination process. There are many stages in current state-of-the-art UXO discrimination strategies, such as phenomenological model inversion, feature generation and selection, and selection of a classification algorithm. If each stage of the discrimination relies on standard, typical methods found in the research literature for model inversion, feature selection, and classifier training, it is reasonable to expect that outliers would be present in the output of each stage since classical approaches are often designed specifically to solve the most common or expected scenario, and accept outliers as part of the distribution of possible observations. To maximize performance of the UXO discrimination strategy at the $P_D = 1$ operating point, and to satisfy the government regulators, it is an interesting basic research question to propose that each stage actively attempts to mitigate the occurrence of outlier UXO. In decision theory, this difference can be distinguished as the difference between error minimization and cost minimization.

Any UXO discrimination strategy is capable of two types of error: making a “dig” declaration for a harmless clutter object (termed a “false alarm” or “false positive”), and making a “no dig” declaration for a UXO (termed a “miss” or “false negative”). As both of these scenarios (false alarms and misses) are errors, an error-minimization approach would be equally motivated to avoid either of these events, since both contribute equally to the overall number of errors. However, a more appropriate consideration in UXO cleanup scenarios is cost, rather than errors. Costs, or penalties, can be assigned to the different combinations of “dig / no dig” declaration and true anomaly class. Since the cost of a miss (not making a “dig” declaration for a UXO) far exceeds the cost of false alarm (unnecessarily making a “dig” declaration for a harmless clutter object), a cost-minimizing decision criterion will shift the bias towards making the least costly errors (i.e. false alarms). This has previously been considered by Carin et al. [8] using a POMDP-based approach to policy implementation. The framework was able to consider various costs, such as the cost of making an observation with another sensor or the cost of making either a “dig” or “no dig” declaration given the data available. However, as will be outlined in the next section, an alternative approach to minimizing the number of dig declarations to reach the $P_D = 1$ operating point may yield better performance when considering potential drawbacks associated with the cost-minimizing approach. Rather than being too dependent on sensitivity to the cost estimates, as are all cost-minimization approaches such as [8], our hypothesis is that we may be able to achieve better performance at the $P_D = 1$ operating point by first analyzing the behavior and system dynamics of the UXO discrimination strategy, examining how outliers are produced and handled at various stages in the data processing, and developing guidelines for processing architecture that will mitigate such events.

General UXO discrimination strategy framework

Many of the current approaches to UXO discrimination currently utilized by the research and cleanup community can be described within the general framework presented in Figure 1. In this general framework, there are four stages in the UXO discrimination process that occur between

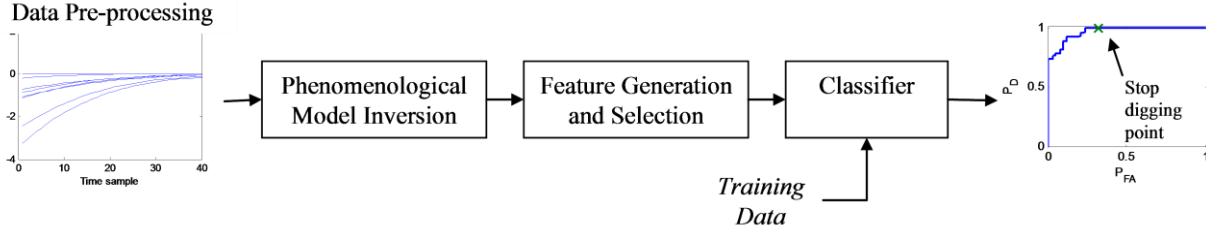


Figure 1. General framework of UXO/clutter discrimination strategy, from raw data to “dig / no dig” declaration. There are four stages: 1) data pre-processing, 2) model inversion, 3) feature generation and selection, and 4) classification using labeled training data.

the measurement of the raw sensor data and making a “dig / no dig” declaration. The first stage is data pre-processing, which includes the selection of sensor measurements and time samples, and application of any normalization or background correction. The second stage of processing involves model inversion to fit the measured sensor data. Models can incorporate various assumptions about the shape of the EMI signatures as a function of time and have different physical or empirical motivations. Often, it is possible to make a trade-off between model rigor and expressivity. Model inversion and numerical optimization is a significant field of research in and of itself. Various methods can be considered to manage the presence of local minima, high-dimensionality of the parameter space, and data quality.

The third stage of processing uses the phenomenological model parameters estimated via the model inversion to generate features for UXO/clutter discrimination. Relevant features can be generated based on knowledge regarding the phenomenological models; ideally, physics suggests that the features should correspond to the intrinsic characteristics of the anomaly in order to be useful for discriminating UXO from clutter. A large set of features, which commonly occurs for more advanced multi-axis sensors, can be reduced through downselection using feature selection techniques. There are many relevant feature selection techniques that can be differentiated based on their methods for generating candidate sets of features and scoring the suitability of different feature combinations. Alternatively, some investigators use the entire model fit and perform library-matching with the polarizability curves. Such an approach fits within the framework presented in Figure 1; the “features” from which decision outputs will be generated are the raw model inversion outputs, generated via a 1-to-1 mapping rather than a more complex feature generation scheme that aggregates the raw outputs into a more concise set of values. The final stage of the UXO discrimination strategy is classification. In this stage, the features associated with an anomaly of interest are analyzed with respect to a set of features from training samples using a classification algorithm, and a score is assigned to the new anomaly by the classifier. Based on various threshold and decision criteria, the score can be converted into a “dig / no dig” declaration. The design of the first two processing stages benefits from a high level understanding of geophysics, sensor phenomenology, and numerical optimization, whereas the design of the final stage benefits from a strong background in statistical signal processing and

pattern recognition. Feature generation and selection requires experience in both areas to be performed in an optimal fashion.

Several groups have successfully implemented various techniques for each stage of the UXO/clutter discrimination framework (e.g. [8-13]). Data pre-processing can include background correction and sensor position uncertainty modeling (e.g. [14, 15]). Model inversion has been performed using standard gradient descent approaches and stochastic, evolutionary inspired methods [16-18]. Published results have used the generalized likelihood ratio test (GLRT), support vector machine (SVM), and artificial neural networks (ANN) as classifier methods in UXO/clutter discrimination. These are standard signal processing solutions to the types of problems posed by each stage of the UXO discrimination strategy. However, it has not been documented whether these techniques are optimized for the operating point required by the UXO cleanup task ($P_D = 1$). Therefore, given the extensive number of parameters and options associated with each stage, it is reasonable to suspect that certain techniques may perform better at the required operating point than others. Additionally, a set of guidelines for use may be beneficial for improving performance at the $P_D = 1$ operating point.

Relevant theory

In this section, the relevant theory and concepts that motivate the methods and techniques applied in this project will be described briefly. There is a relevant body of work from the decision theory, pattern recognition, and machine learning literature that can be leveraged in this effort to maximize the performance of a UXO discrimination strategy designed to perform at a specific operating point. This section will also highlight the novelty of the work performed in this investigation.

Statistical decision theory offers several methods for determining the operating point of the UXO/clutter discrimination strategy, such as the Bayesian approach for calculating the expected cost of each decision or the Neyman Pearson criterion that optimizes performance at a user-defined number of false alarms or missed detections. However, the drawback of any of these criteria associated with decision theory is that they are applied *ex post*, i.e. once the classifier outputs are already determined. These approaches find the desired operating point on the ROC curve, but do not explicitly improve performance at that point. To improve performance, the ROC curve which describes all possible operating points for the classification system needs to also improve. As is well established, there can be substantial variability within the ROC curves produced by a classifier on a fixed set of testing and training data, simply by changing the parameters associated with most classifiers. Beran and Oldenburg observed such trends in comparisons of a support vector machine, a probabilistic neural network, and library based techniques on data sets from two sensors, and recommended a careful approach to the design of a classification method for UXO discrimination [19]. In the UXO discrimination task, we are willing to sacrifice performance at operating points not of interest in UXO/clutter discrimination to improve performance at $P_D = 1$. Therefore, what is proposed in this research is

a systematic methodology to determine what characteristics of the classifier might produce ROC curves with better performance at the $P_D = 1$ operating point. Decision criteria, either cost minimization or Neyman Pearson, can help find that point on the ROC, but they alone cannot improve performance at the $P_D = 1$ operating point in the classifier design stage.

A body of work has focused on the area of cost-sensitive classification. Recognizing the distinction between minimizing error and minimizing cost, other researchers have sought to modify error-minimizing classifiers, such as the decision tree and support vector machine, to operate in cost-minimization scenarios. Two dominant approaches have emerged from this research area. The first method is termed stratification [20]. In stratification, the set of training data is modified, either through weighting or resampling of the data points, such that the proportion of samples from each class is consistent with the costs. To implement such a technique in the UXO discrimination task would require significant modification to the design of the training data set. For the UXO cleanup task, labeled samples for training the classifier are already of limited availability and difficult to collect; thus such extensive modification to the training set is difficult to support. Additionally, given the highly disproportionate costs of missed UXO and false alarms, it is anticipated that the stratification framework would simply reject all clutter samples and retain only UXO. The second technique for cost-sensitive classification is MetaCost [21], a wrapper method that extends cost-sensitive classification to any classification algorithm. In MetaCost, the training samples might be relabeled according to their “cost-minimizing label” if it differs from the true training sample label. The classifier is then re-trained using the cost-minimizing class labels for the training data. These cost-minimization based techniques (stratification, MetaCost, and the previously-mentioned POMDP-based approach to policy implementation) may not yield ideal results due to their dependency on the estimated costs for the possible errors. The elegant simplicity of Bayesian cost-minimization may hamper its use in UXO discrimination, since $P_D = 1$ can only be satisfied theoretically if the cost of a missed UXO is asymptotically approaching infinity, and “dig” declarations are made for all anomalies. Even when provided with more realistic estimates of cost, the decisions will be quite sensitive to the estimates of cost, resulting in potentially unstable performance near the operating point.

A recent trend in the development of pattern classification methods is the increasing popularity of hybrid and meta-classifiers, which combine two or more component classifiers in an overall classification scheme. These methods have been shown in many instances to provide improved performance and robust behavior in several applications [22-24]. There are various design considerations in the construction of a meta-classifier: the training data used for each component, the selection of component classifiers, and the structure of the meta-classifier. Training data can be sampled randomly using a technique termed bootstrapping, or specifically tailored for each component to accomplish a specific task (i.e. context-dependent training). Component classifiers can all be of the same type, such as a collection of classification trees used in the Random Forest classifier [25], or they can have distinct characteristics, as in the

generative/discriminative hybrid classifier frameworks [22-24]. Finally, the meta-classifier can be structured as a parallel combination of components, as in the Random Forest or hierarchical mixture of experts [26], or cascaded, as in the popular AdaBoost method [27] and several implementations of hybrid classifiers [23, 24].

Materials and Methods

The experiments conducted in this SEED effort relied extensively on data collected as part of the Former Camp San Luis Obispo demonstration. Data and model fits were provided by Snyder Geoscience for the MetalMapper sensors with the MM/RMP model [28]. This model performs sequential stages of estimation: first, nonlinear estimation to determine position and the symmetric matrix of polarizability transients, then linear estimation of the three attitude angles and three principal polarizability transients, followed by parametric curve-fitting. A total of 1072 anomalies (887 clutter, 185 UXO) were taken from the cued identification survey and used for the discrimination experiments in this study. From the set of available model parameters, a subset of nine features were selected; thus, the data was capable of being formatted in a 1072 by 9 matrix. The motivation for manual selection of a feature subset rather than using empirical results from feature selection algorithms was to reduce the opportunity for bias towards one of the classification algorithms that was to be evaluated in this study. The most common methods for feature selection (either information-theoretic filter methods or classifier-dependent wrapper methods [29, 30]) will assume an underlying model for the distribution of anomalies in each class. Therefore, the feature subset was selected based on descriptions of the physical target characteristics that are represented by each feature and their likelihood for representing intrinsic characteristics of the anomaly. Based on this analysis, the following nine features were selected:

- *NormP0, NormP1*: RMS values of $P0X, P0Y, P0Z$ and $P1X, P1Y, P1Z$, which represent the numerical integration of the principal polarizability curves and their first moments, respectively.
- *P1X, P1Y, P1Z*: Numerical integration of the first moment of the principal polarizability curves.
- *P0_R, P0_E*: Measure of aspect ratio and eccentricity based on principal polarizability curves.
- *P1_R, P1_E*: Measure of aspect ratio and eccentricity based on first moment of the principal polarizability curves.

Through follow-up communications with Skip Snyder of Snyder Geoscience it was revealed that six of the nine features match up with the feature set they used at the recent Camp Butner demonstration¹. To illustrate the separation of UXO and clutter objects using the nine features selected for use in this study, a two-dimensional principal component projection of the feature

¹ Snyder used $P1_T$, equal to the average of $P1Y$ and $P1Z$, in addition to $P1X, P0_E, P0_R, P1_E$, and $P1_R$.

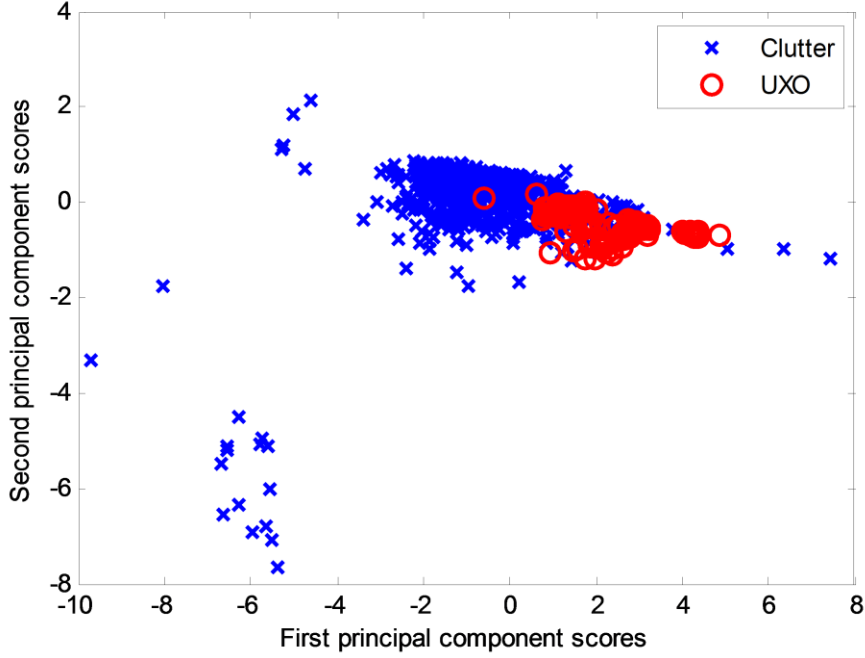


Figure 2. Representation of the nine-dimensional feature space in two dimensions using principal components analysis. The x- and y-axis correspond to the scores along the first and second principal components, respectively.

set is shown in Figure 2. The lower-dimensional subspace reveals strong clustering of the UXO, with greater variability observed in the clutter features

This SEED project relied extensively on the use of large-scale Monte Carlo simulations to thoroughly analyze the properties and performance of the different algorithms and stages in the UXO discrimination scheme. Large sets of independent classification results were generated by randomly-sampling from the available set of 1072 anomalies. Two methods of sampling were employed in this study. The primary means of constructing subsets of data was by sampling without replacement to divide the 1072 anomalies into two data sets (one for classifier training and a second for validation). Alternatively, for certain experiments it was desired to measure the distribution of a measure or statistics (e.g. $P_{FA} @ P_D = 1$). In these instances, bootstrapping (sampling with replacement) was utilized to estimate the underlying probability distribution for the measure of interest.

One of the main focuses of this SEED study was the statistical classification algorithm that is used in the final stages of the UXO/clutter discrimination strategy. There are many possible choices for algorithms to use in this stage, and several experiments examined the various properties of difference classifiers to evaluate their potential for operating at $P_D = 1$. Table 1 lists the classifiers considered in this study, along with several significant properties. The third column of the table indicates whether the classifier is generative or discriminative.

Generative classifiers attempt to model the distributions of each class (i.e. UXO and clutter), and use that information to make classification decisions. Discriminative classifiers do not attempt to learn the class-specific feature distributions and instead only attempt to learn the boundary line separating the two classes. Thus, the discriminative classifier is often viewed as having a simpler learning task. The second classifier property is local versus aggregate use of training data. Local classifiers make a classification decision based only on the neighboring training samples, whereas aggregate classifiers rely on parameters that are calculated from all of the available training data. A simple test for whether a classifier uses “local” or “aggregate” training data can be conducted by analyzing whether the classifier’s output for some test sample x_{TEST} would be sensitive to the addition of a large amount of new training data at a point in feature space not near x_{TEST} . If the classifier’s output is not affected by the addition of new training samples, the classifier makes “local” decisions. The final classifier property specified in the table is parametric versus nonparametric classifiers. Parametric classifiers make use of a model to condense the information in the training data to a finite number of parameters, whereas a nonparametric classifier preserves the entire set of training data for making decisions on test data. Thus, the storage requirements increase for nonparametric classifiers as more training data is acquired.

Table 1. List of classifier algorithms and relevant properties.

Classifier	Acronym	Generative / Discriminative	Local / Aggregate	Parametric / Nonparametric	Reference
Generalized Likelihood Ratio Test	GLRT	Generative	Aggregate	Parametric	[31]
Distance Likelihood Ratio Test	DLRT	Generative	Local	Nonparametric	[32]
K Nearest Neighbor	KNN	Generative	Local	Nonparametric	[33]
Linear Discriminant	FLD	Discriminative	Aggregate	Parametric	[33]
Support Vector Machine	SVM	Discriminative	Aggregate	Parametric	[26, 34]
Relevance Vector Machine	RVM	Discriminative	Aggregate	Parametric	[35]
Random Forest	RF	Discriminative	Aggregate	Nonparametric	[25]
Artificial Neural Network	ANN	Discriminative	Aggregate	Parametric	[33]

Results and Discussion

Task #1: Classifier design

The first experiments in Task #1 focused on assessment of the performance of the different classifiers (listed in Table 1) at $P_D = 1$. The performance of these classifiers was analyzed based on performance statistics for 1000 data sets generated using the Monte Carlo sample-without-replacement experiment design. Each of the 1000 test sets consisted of 100 randomly-selected anomalies; the remaining 972 anomalies were used for training data. The P_{FA} at $P_D = 1$ was calculated from the resulting ROC curve for each data set and each classifier. Figure 3 shows the distribution of P_{FA} values over the 1000 data sets. In these box plots, the red line indicates the median value, the edges of the box correspond to the 25th and 75th percentile, and the whiskers show the extent of the remaining data points (excluding outliers, shown as red markers, which are above the 75th percentile or below the 25th percentile by more than 1.5 times the interquartile difference). The DLRT and Random Forest classifiers appear to have most of the results with the lowest P_{FA} values. The FLD and SVM classifiers were able to achieve low P_{FA} for many of the data sets, but also have a large proportion of instances where P_{FA} is quite high (greater than 0.8). The GLRT classifier has a similarly large range of P_{FA} values, and seems to be quite dependent on the specific training/test data set.

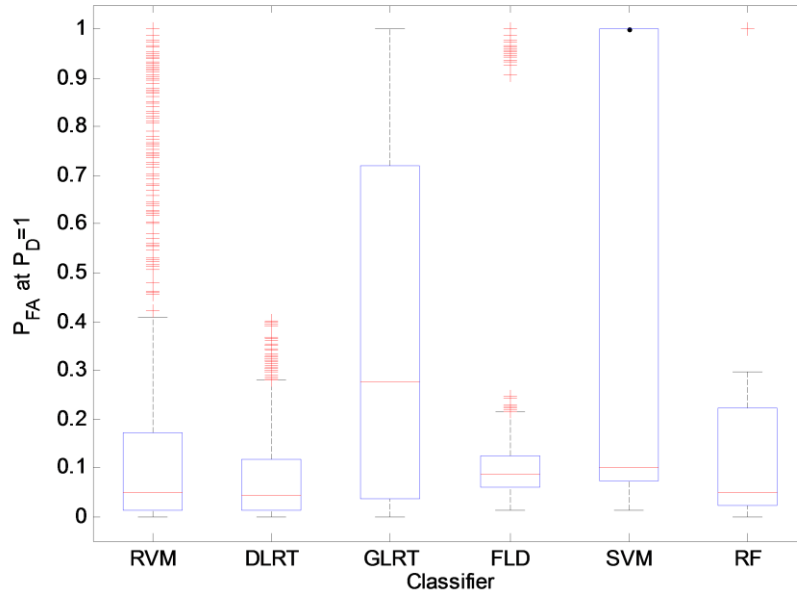


Figure 3. Boxplot showing distribution of P_{FA} at $P_D = 1$ over 1000 sets of training/test data for six classifiers. Red lines identify the median of the distribution and the edges of the blue box extend to the 25th and 75th percentiles. The red hash marks identify outliers that are beyond the 25th and 75th percentiles by more than 1.5 times the interquartile distance.

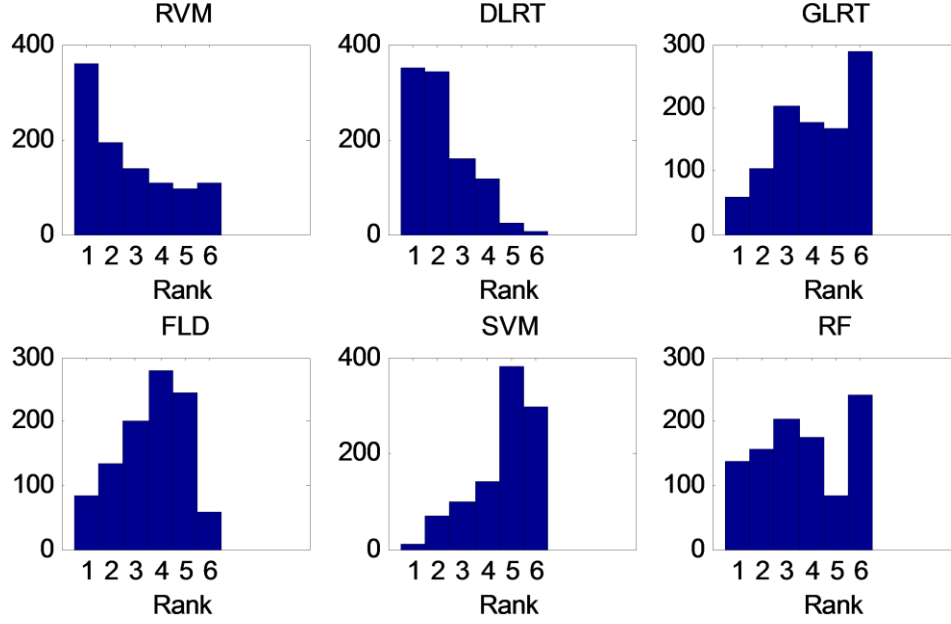


Figure 4. Performance rankings for each classifier algorithm over all 1000 training/test sets corresponding to the results presented in Figure 3. Each column in each subplot indicates the number of instances (out of 1000 possible) that the classifier had the lowest P_{FA} (rank of 1) though highest P_{FA} (rank of 6).

An alternative view of these experiments was generated by ranking the P_{FA} values for the six classification algorithms (from 1 to 6, with 1 corresponding to the lowest P_{FA} value) for each of the 1000 test cases. A histogram of the rankings for each classifier is shown in Figure 4. In this analysis based on the ranking of each classifier (in terms of P_{FA}), the DLRT and RVM classifiers appear to have the best performance; i.e. for most data sets they outperform the other four classifiers and provide either the lowest or second-lowest P_{FA} .

A second experiment was set up to examine the propensity of a classifier to declare UXO at points in feature space where it has previously observed clutter in the training dataset. This classifier property may be necessary to extend the decision boundary to encompass enough of the feature space to operate at $P_D = 1$. In this experiment, each classifier was trained on all available data, except for three randomly-selected UXO held out for use in the test set. In the place of these three UXO, three observations labeled as “clutter” were inserted into the training data with the same feature values as the held-out UXO. In the test stage, the classifier was evaluated on the three held-out UXO (with feature values matching the “clutter”-labeled observations inserted into the training data) and the clutter from the training set. The performance metric calculated for each of the 1000 iterations in this experiment was P_{FA} with 100% detection of the 3 hold-out UXO. The rankings (by lowest P_{FA}) across the 1000 iterations are shown in Figure 5. The DLRT classifier was capable of detecting the three UXO with the lowest P_{FA} for most datasets. This suggests that the DLRT is the classifier most capable of ignoring a few clutter at the periphery of the main UXO cluster in feature space to find UXO that are mild outliers.

Another comparison was made to examine the similarity of the decisions produced by each classifier. If two classifiers produce decisions that are not highly correlated, then the combination of the classifiers may result in a system that has greater information available for detecting UXO, leading to higher performance with the fusion-classifier approach. In this experiment, a test set was constructed by randomly drawing 10 UXO and 10 clutter items. This test set was held fixed for 100 iterations of the classification study using training data that consisted of 150 anomalies from each class (300 in total) drawn at random. Thus, a matrix of decision metrics with dimensions (nClassifiers x 100) by 20 could be constructed (where nClassifiers is the number of classifier algorithms considered in the experiment). The decision metrics were normalized within each row of the matrix to be unit-variance and zero mean, and the distances between the decision outputs for each iteration (i.e. each different training set) and each classifier were calculated. This distance matrix can then be decomposed using principal component analysis and represented in a two-dimensional space. This technique is known as multidimensional scaling [33], which attempts to represent a data set in a lower-dimensional space while maintaining the relative distances between points in the higher-dimensional space. If any individual classifier is not affected by the variability in the training set over the 100 iterations, then the decision metrics would be identical for that classifier and the calculated distances would be zero, thus all 100 points for that classifier would occur at a single point in the multidimensional scaling subspace.

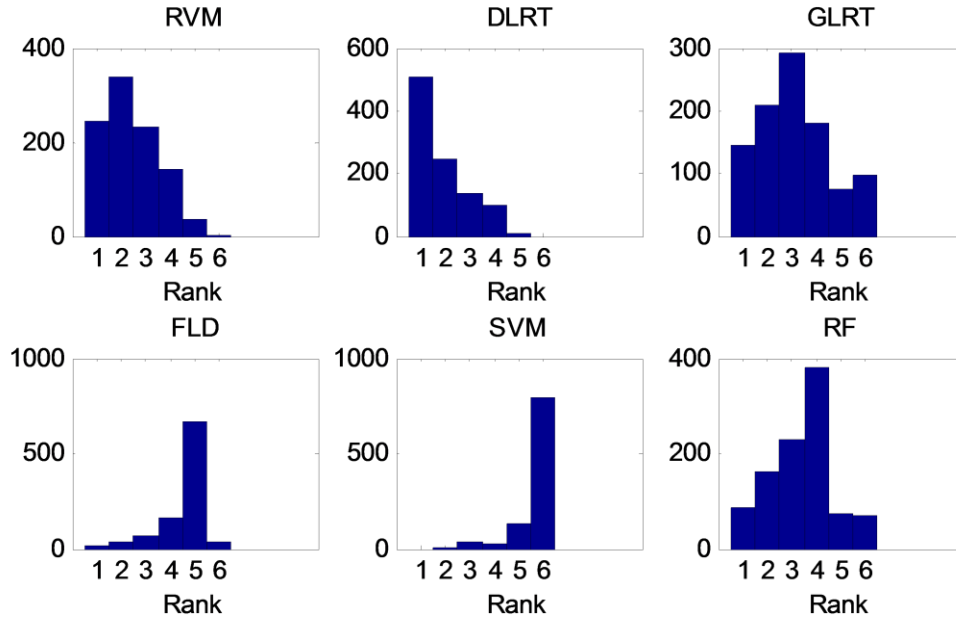


Figure 5. Performance rankings for each classifier algorithm over all 1000 training/test sets in the experiment testing classifier propensity to declare UXO. Each column in each subplot indicates the number of instances (out of 1000 possible) that the classifier had the lowest P_{FA} (rank of 1) though highest P_{FA} (rank of 6).

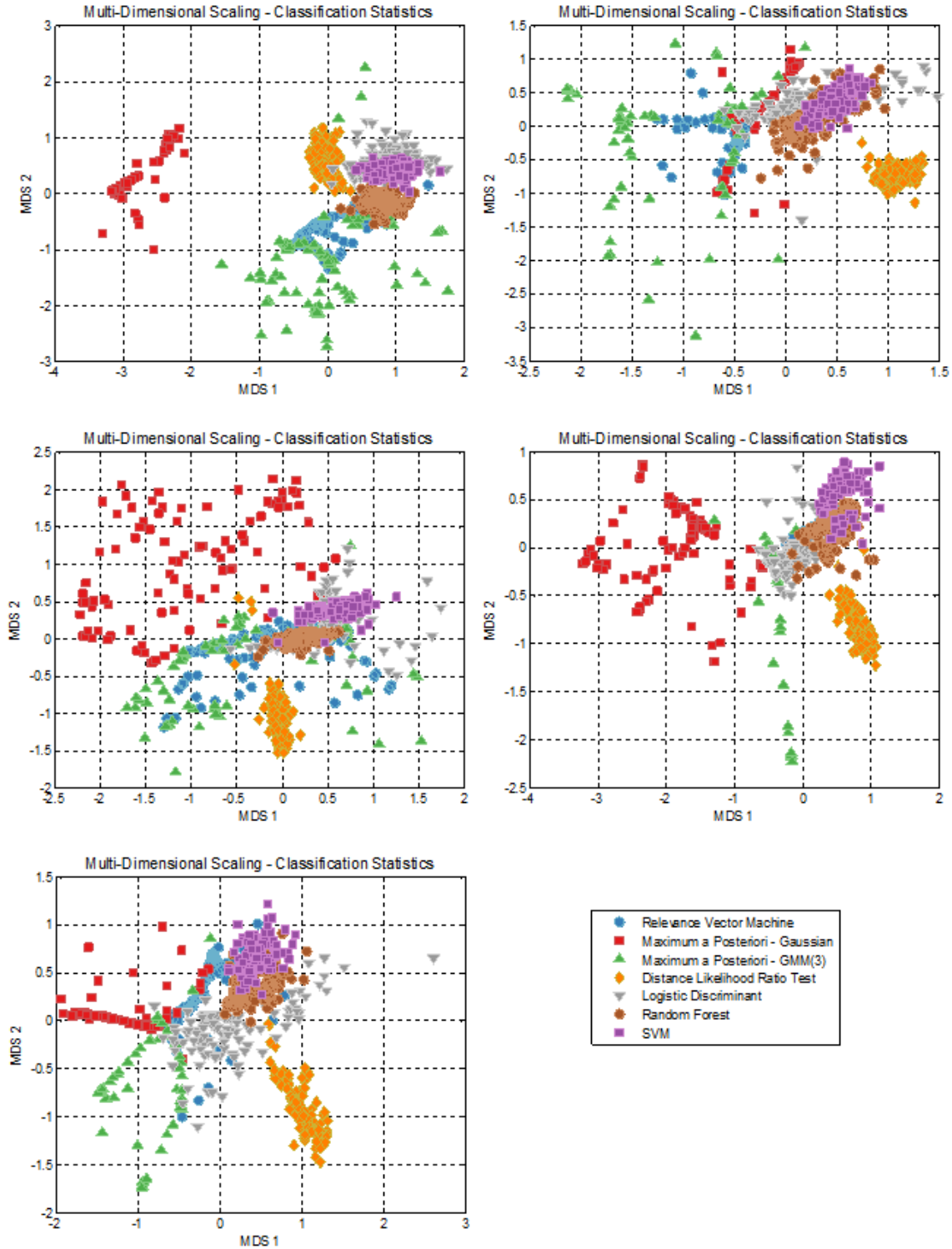


Figure 6. Multidimensional scaling representation of the classifier algorithm outputs for five distinct test sets (corresponding to the five subplots). Within each subplot, each point represents the set of decision metrics that result when using a unique set of training data. Different marker types correspond to the different algorithms. Points that are close in proximity represent similar decision patterns, whereas points located far apart represent dissimilar classifier decision outputs.

The multidimensional scaling analysis was run for five different randomly-selected test data sets. Each two-dimensional projection is shown in Figure 6. Across the five test sets, there are some common trends. The DLRT and SVM tend to form fairly tight groups; thus, their decision metrics are fairly consistent despite the variability across the 100 instances of the training data set. The Random Forest and SVM classifiers are in close proximity in the MDS space suggesting that these two classifiers produce similar decision outputs. The MAP classifier and the DLRT often appear on opposite sides of the MDS subspace, which indicates that these two classifiers may produce some of the most different decision outputs. The plot also reveals that the MAP classifiers are particularly sensitive to the choice of training samples, based on the spread of the points for these classifiers.

Another element in Task #1 was assessment of the effects of training data. The first experiment, which performed a baseline evaluation of performance (P_{FA} at $P_D = 1$), was re-run using significantly smaller training sets. In this experiment, 100 anomalies were randomly selected for use in training (with a minimum of 20 UXO) and the remaining 972 anomalies were

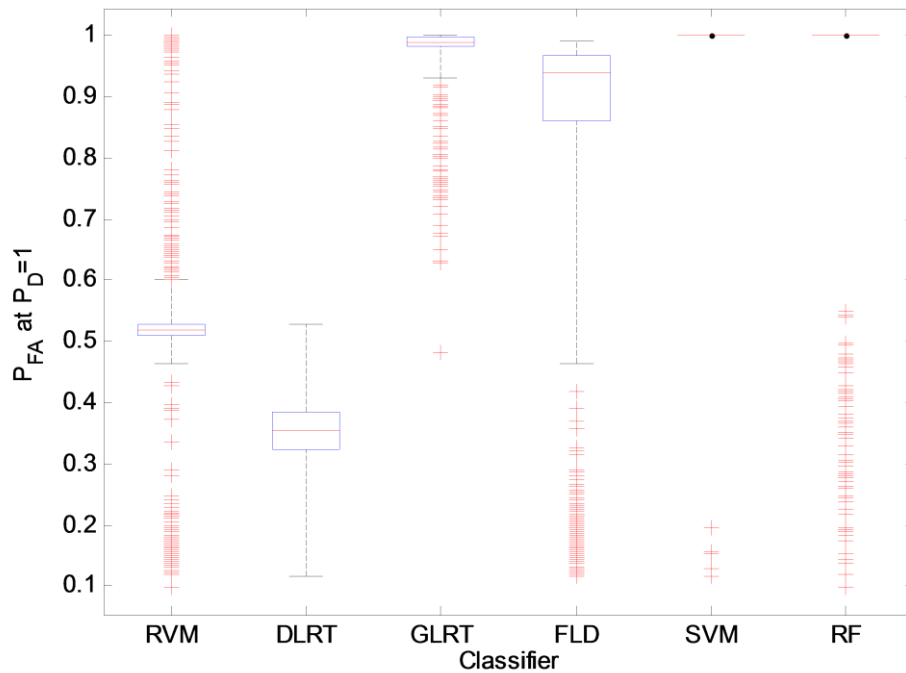


Figure 7. Boxplot showing distribution of P_{FA} at $P_D = 1$ over 1000 sets of training/test data for six classifiers when the amount of available training data is significantly reduced (100 anomalies, at least 20 of which are UXO). Red lines identify the median of the distribution and the edges of the blue box extend to the 25th and 75th percentiles. The red hash marks identify outliers that are beyond the 25th and 75th percentiles by more than 1.5 times the interquartile distance.

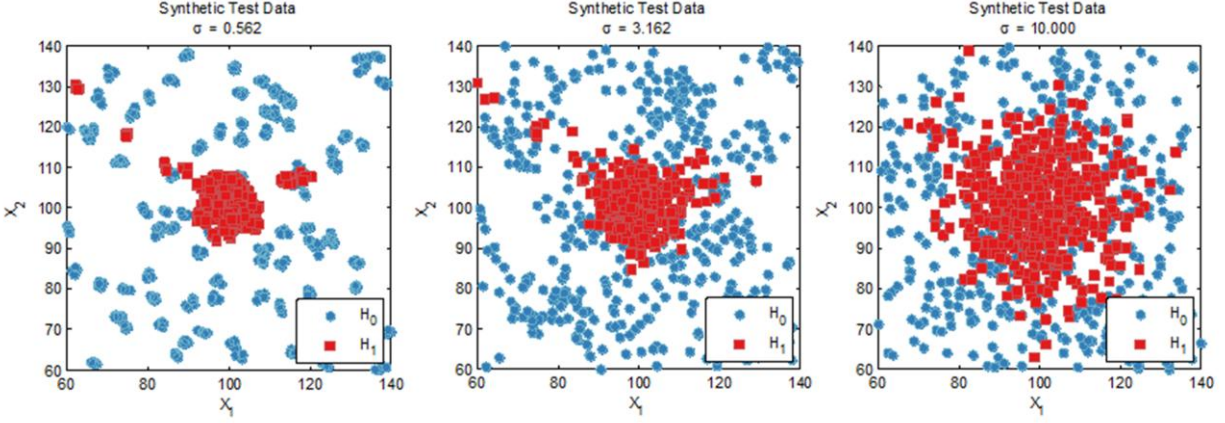


Figure 8. Illustration of the variability in the synthetic data set introduced by increasing the noise σ_n .

used at the test set. A total of 1000 iterations were run in this experiment. The distribution of P_{FA} at $P_D = 1$ is represented in Figure 7 as a boxplot. In this scenario, the significantly reduced sized of the training data set results in much greater variability in P_{FA} at $P_D = 1$. However, the overall pattern of performance is little changed; the DLRT consistently is the best performing classifier in this reduced training data scenario.

Another consideration is the effect of mismatch between the training and test data. An experiment was run using synthetic data with mismatch between the training and test data was introduced by adding Gaussian noise to the test set features. Examples of the test data for three different noise levels are shown in Figure 8. In this experiment, the test data consisted of 1000 anomalies equally split between H_1 and H_0 samples. Three training set sizes were investigated ($N=200, 400$, and 1000) with equal representation of H_1 and H_0 . A total of 800 iterations were run to generate distributions of P_{FA} at $P_D = 1$.

The distribution of P_{FA} at $P_D = 1$ for each classifier was represented using two statistics: the average and standard deviation of P_{FA} over the 800 iterations. These values can be represented as a point in a 2-D plot, as show in Figure 9 *left* for all classifiers. The multiple points connected by a line represent the increasing mismatch between the training and test sets due to increasing σ_n . The ideal performance point is at $(0,0)$ in the lower left corner. When both performance and predictability are considered, the $DLRT_1$ and GMM GLRT appear to provide the best results. It is worth noting that the results are fairly sensitive to some classifier parameters. The DLRT classifier was run with three values of K , the sole user-specified classifier parameter, which determines the size of the neighborhood used to classify a new test sample. This sensitivity is further highlighted in Figure 9 *right*, which shows P_{FA} at $P_D = 1$ for an extensive range of values for K , measured using 10-fold cross-validation on the entire 1072 anomaly data set. Figure 9 *right* reveals that while the DLRT classifier is capable of performing well at $P_D = 1$, it is sensitive to the parameter K .

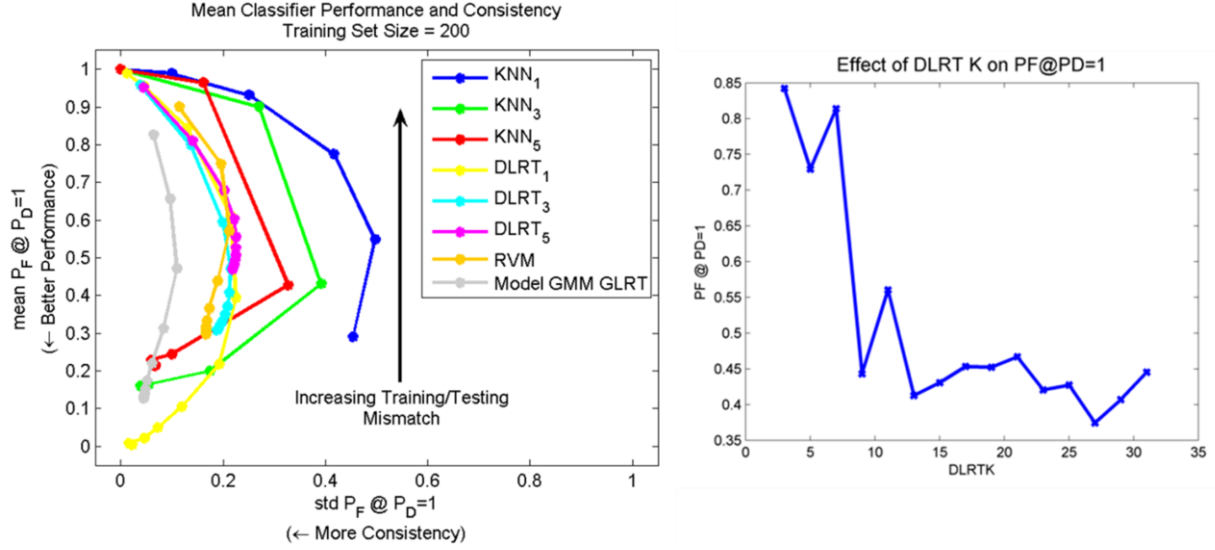


Figure 9. *Left:* Effects of mismatch in the training and test data, represented by the mean P_{FA} at $P_D=1$ (y-axis) and standard deviation of P_{FA} at $P_D=1$ (x-axis). P_{FA} at $P_D=1$ statistics are calculated from 800 training/test data sets, and each point represents a distinct mismatch condition (i.e. amount of noise added to produce mismatch between training and test data). *Right:* Effect of DLRT parameter K on PF at $P_D=1$, as measured based on 10-fold cross-validation use the baseline data.

In a similar format to Figure 9, Figure 10 shows the results for the same training set mismatch study using trainings sets with two larger sizes: $N = 500$ (left) and $N = 1000$ (right). One trend that can be observed as the size of the training data increases is the reduction in variability of the results for the DLRT classifier, suggested by the leftward shift of the three lines representing the DLRT. Thus, the standard deviation of P_{FA} at $P_D=1$ for these classifiers is decreasing as the amount of training data increases. This result is supported by previous studies that have shown the benefit of larger training set sizes for nonparametric classifiers (e.g. [36]).

Task #2: Sensitivity to feature selection and model inversion

In Task #2, the focus of the project shifted to the prerequisite stages in the UXO discrimination strategy: model inversion and feature selection. Two studies of feature selection were motivated by questions about the impact of the size of the feature set on performance at $P_D = 1$ as well as the most appropriate performance measure for feature selection based on empirical feature selection methods.

In the first study, feature sets of various sizes were considered with four different classifiers. The feature sets were organized into three characteristic groups: two time features (Tm), four size features (Sz), and six shape features (Sh). The features in this experiment were selected after consultation with Skip Snyder at Snyder Geoscience. The feature groups were evaluated individually and in combination, which produced a 12-feature superset in the latter case. The experiment used 10 realizations of 10-fold cross-validation, producing 100 calculations

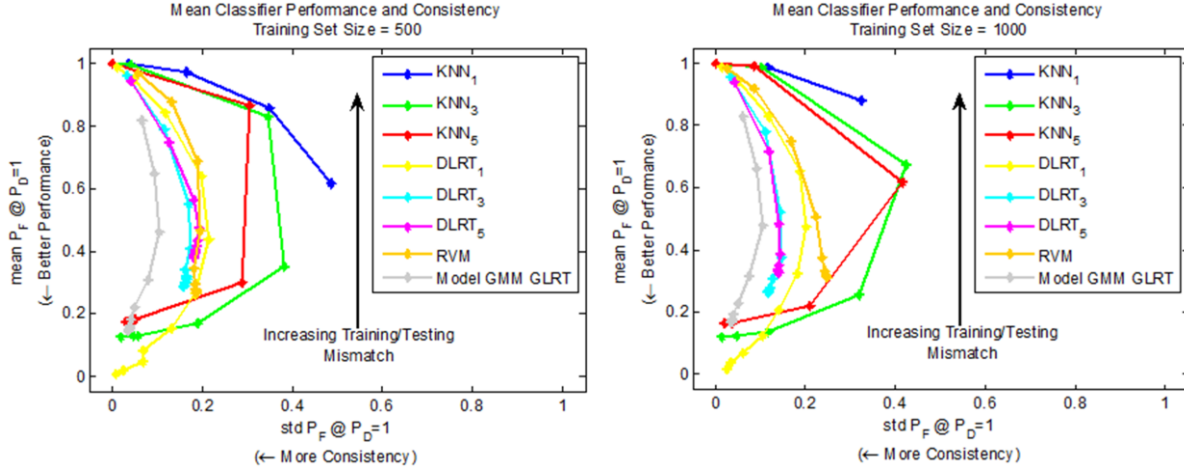


Figure 10. Effects of mismatch between training and test data, expressed using statistics of distribution of P_{FA} at $P_D = 1$, when the training set size is increased to $N = 500$ (left) and $N=1000$ (right).

of PF at $P_D = 1$. Figure 11 shows the distribution of P_{FA} at $P_D = 1$ for the various individual, paired, and complete feature set for each of the four classifiers. In the subplots, the number of features used for classification increases from the bottom (T_m) to the top ($T_m + Sh + Sz$). Based on inspection of the median of the P_{FA} distributions (i.e. the red line in each boxplot) for each feature set, there does not appear to be a strong trend in performance as a function of feature set. This result is not consistent with conclusions from an earlier SERDP-sponsored project (MM-1442), where it was observed that larger feature sets provided better performance. One possible explanation for the inconsistency is that the previous study used empirical feature selection techniques to sequentially determine the feature sets.

A second study examined various possible performance measures for empirical evaluation of a feature set / classifier combination. The most obvious measure is P_{FA} at $P_D = 1$ since this is the desired operating point for the designed system. However, it may also be useful to measure P_{FA} at $P_D = \left(1 - \frac{1}{N_{UXO}}\right)$ and $P_D = \left(1 - \frac{2}{N_{UXO}}\right)$, where N_{UXO} is the number of UXO in the test set. These are performance measures that correspond to nearly-perfect UXO detection by our system. The area under the ROC curve (AUC) is a commonly used measure, as ROC curves have often been used to present the results of UXO discrimination studies. More relevant to our desired operating point is AUC for $P_D > 0.95$, which calculates just the area under the upper part of the ROC curve. Finally, the equal error rate (EER) was also included in this investigation. The equal error rate is the value ε such that $(1 - P_D) = P_{FA} = \varepsilon$, and it is a commonly-used performance measure in other fields.

In this experiment, the sampling-without-replacement method was used to generate test and validation data sets, each with 50 anomalies. The remaining 972 anomalies were used as training data in the DLRT classifier. The experiment was repeated 1000 times, with each of the performance measures described above calculated for both the test and validation set. The

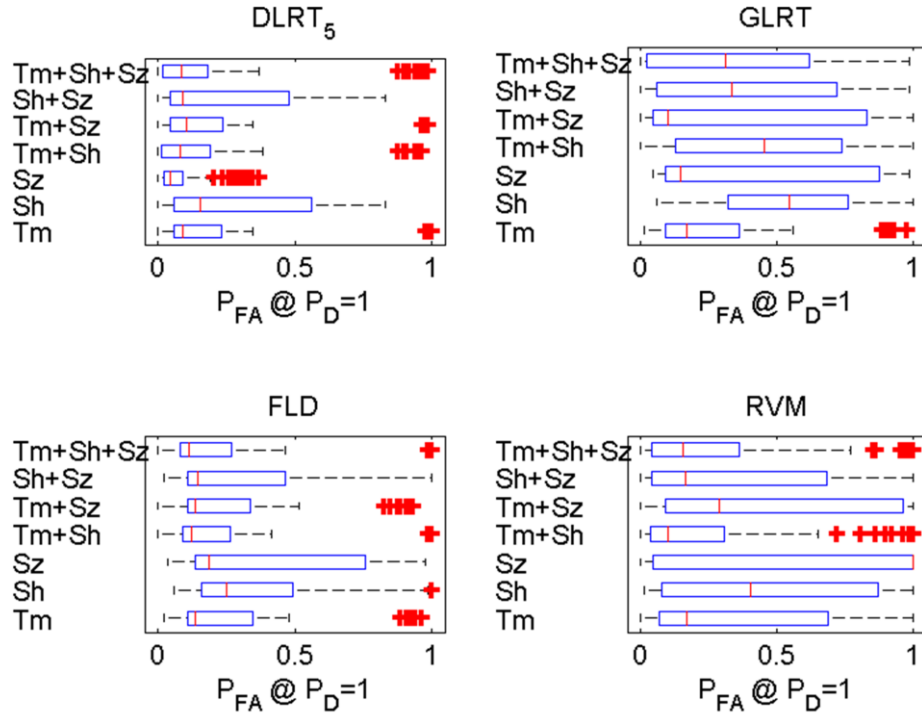


Figure 11. Boxplots showing distribution of P_{FA} at $P_D = 1$ using seven feature sets of increasing dimensionality with four different classifiers. P_{FA} distributions were calculated from 100 training/test data sets (10 repeats of 10-fold cross-validation). Red lines identify the median of the distribution and the edges of the blue box extend to the 25th and 75th percentiles. The red hash marks identify outliers that are beyond the 25th and 75th percentiles by more than 1.5 times the interquartile distance.

results generated a 1000 by 12 matrix (six performance measures for both test and validation data sets). Figure 12 shows the correlation matrix, i.e., the correlation between the performance measures over the 1000 iterations of the experiment. The distinct blocks of pixels in the upper left and lower right of the image correspond to correlations of measures *within* the same data set. The blocks of pixels in the upper right and lower left correspond to correlations of performance measures *across* the two data sets.

Within a set, the AUC for $P_D > 0.95$ is most correlated with our desired operating point (P_{FA} at $P_D=1$). Somewhat surprisingly, there is very little correlation between P_{FA} at $P_D = 1$ and P_{FA} at the other P_D levels considered: $P_D = \left(1 - \frac{1}{N_{UXO}}\right)$ and $P_D = \left(1 - \frac{2}{N_{UXO}}\right)$. Thus, what may be assessed as a good-performing feature set/classifier combination at one measure may change if an incrementally-lower P_D is considered for the system. Looking to the between-set results, there is very little correlation between any of the measures calculated on the separate data sets. Ideally, high correlation between P_{FA} at $P_D = 1$ in Set 2 and any performance measure in Set 1 would indicate good predictive power of estimating future performance on new data. However,

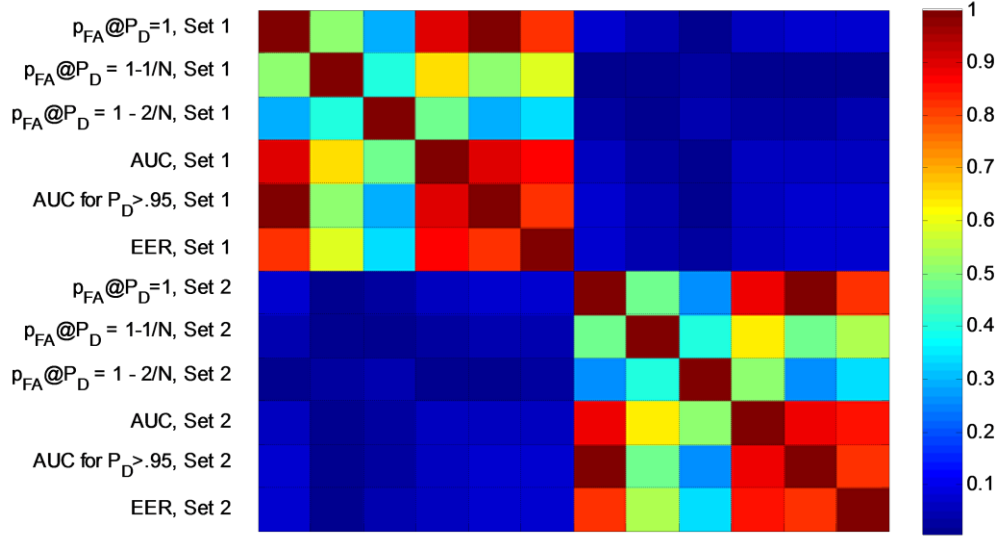


Figure 12. Matrix of correlations between various performance measures calculated within and across two independent test sets.

the performance measures evaluated in this study appear to provide very little information across data sets.

The model inversion stage of the UXO discrimination strategy is one place where mitigating the occurrence of UXO outliers is critical to improving performance at $P_D=1$. An experiment was performed to investigate a potential technique for identifying outliers generated by the model inversion. This technique is based on the hypothesis that the model inversion process is fairly robust and consistent. If two data files X_1 and X_2 are very similar, even identical, then the corresponding feature sets that result from model inversion should also be very similar. Note that the inverse relationship does not need to hold true: due to extrinsic parameters in the model and different object orientations, two UXO with very different measurements may correctly have very similar feature representations.

To implement this outlier measure, two matrices of correlation coefficients (R_D and R_F) were calculated. The (i^{th}, j^{th}) entry in R_D contains the correlation between the measurements for the i^{th} and j^{th} anomaly. Similarly, the (i^{th}, j^{th}) entry in R_F contains the correlation between the feature representations of the i^{th} and j^{th} anomaly (thus, both matrices are symmetric). The outlier measure for the i^{th} anomaly is calculated based on the following equation:

$$A(i) = \frac{1}{N} \sum_j 1_{\{R_D(i,j) > R_F(i,j)\}}$$

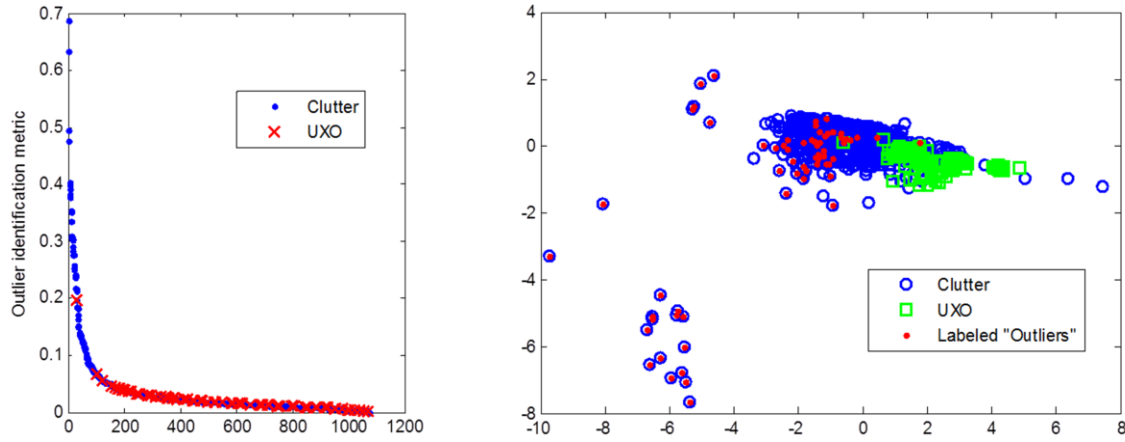


Figure 13. *Left*: Scatter plot of the outlier identification metric A for each anomaly. Marker color and style differentiates UXO and clutter. Higher values of the outlier identification metric indicate a greater likelihood that an anomaly is an outlier. *Right*: Scatter plot of the features in a two-dimensional space (i.e. Figure 2) with potential outliers identified via the outlier identification metric. Red points identify anomalies with an outlier identification metric greater than 0.1 (6% of the total anomaly set). A single UXO is identified as an outlier, and it is in fact the UXO that is most separate from the main cluster of UXO in feature space.

where $1_{\{ \}$ is the indicator function, equal to one if $R_D(i, j) > R_F(i, j)$ and zero otherwise (MATLAB code available in the Appendices). The outlier measure has the effect of looking for dissimilarities between clusters in data-space and clusters in feature-space. When the outlier metric $A(i)$ is large, it indicates that the i^{th} anomaly is dissimilar (i.e. low correlation $R_F(i, j)$) to a number of anomalies in feature space with whom it was more similar (i.e. higher correlation $R_D(i, j)$) in data space.

The values of the outlier metric for the 1072 anomalies in the data set are plotted in Figure 13 *left*. Most of the anomalies have a low value of the outlier metric, indicating that the correlation in data space rarely exceeds correlation in feature space. Selecting 0.1 as an arbitrary threshold, 64 anomalies (6%) are above this threshold and could be further investigated as outliers. It is worth noting that this set of flagged anomalies contains one UXO which has a significantly higher outlier metric than any other UXO. Looking at the two-dimensional representation of the feature space plot (Figure 13 *right*), it can be seen that the identified UXO is indeed the most significant UXO outlier². Additionally, several of the outlier clutter anomalies are also identified.

Conclusions and Implications for Future Research

Recent SERDP/ESTCP projects have focused on the development of sensors that are capable of consistently producing high-quality data. However, high-quality data cannot be fully utilized

² This UXO anomaly is Master ID 1475, a 2.36 inch motor - fins, 60mm boom

without the appropriate follow-on signal processing and analysis. With the goal of operating efficiently at $P_D=1$, this project focused efforts on three stages present in most UXO discrimination strategies: classifier design, feature selection, and model inversion.

Results from the various simulations conducted in this project suggest that careful consideration of $P_D=1$ should be included as part of the algorithm selection. This study provides evidence that the desire to operate at $P_D=1$ may lead to a preference for certain algorithms in the stages of the discrimination strategy. In classifier design, the DLRT classifier consistently performed well. As a nonparametric, neighborhood-based classifier, the DLRT should be particularly appropriate for the $P_D=1$ operating point and the types of feature space that might be observed in UXO discrimination. The numerous simulations and analysis also revealed different behaviors by different types of classifiers when faced with UXO identical to anomalies observed previously, the effects of training set size, and the ability to extrapolate when training data is mismatched (or sparse). Due to the complexity of operating at $P_D = 1$, it is unlikely that a single classifier will be sufficient, and a multistage classification approach that fuses outputs and decisions from multiple algorithms is most likely necessary for robust performance. Evidence supporting such an approach can be found in the results presented in this report. This study also revealed the difficulty in estimating future performance on new data sets. The performance measures most commonly used for empirical selection of feature sets exhibited very little correlation across test and validation data sets independently drawn from the master set of anomalies. Thus, estimating the potential performance of a feature set /classifier combination at $P_D = 1$ has proven to be a difficult challenge.

The investigations of sensitivity to feature set size did not produce conclusive results nor yield any particular recommendations for either large or small feature set size. However, the investigation did reveal avenues of research that could be further explored in the future. More sophisticated sensors and models have the tendency to produce larger feature sets, making the decision about how many features to use (e.g. 6 versus 26) a more relevant question. Additionally, with the use of library-matching techniques for classification and use of the polarizability curves as *de facto* “features”, a further investigation of the effect of statistical pattern classification techniques applied to small and large feature sets for operation at $P_D = 1$ should be considered. The previously-mentioned suggestion regarding classifier fusion should also incorporate an investigation of the most appropriate feature sets to be used in the different component classifiers.

The model inversion procedure is a critical stage in the data processing, and a likely potential source of outliers. The results of the experiments in this study suggest that analyzing model inversion results in the context of all available anomalies may be useful for identifying aberrant model inversions. Because of the importance of the model inversion procedure for successful operation at $P_D = 1$, it will likely benefit from additional analyses such as those investigated in this study to further assess the model fits. Also, the presence of outliers produced by the RbstMultiPrince method, and the ability to detect them based on this divergence of

data/feature correlation, suggest the potential for improvements to model inversion for operating at $P_D = 1$.

Several of the results in this SEED project encourage further investigation; most significantly, the classifier characteristics observed over the various sets of experiments and the potential seen in the model inversion approach. Given the uniqueness of the various classifiers considered (Figure 6) and the performance characteristics in different conditions of training/test data (Figure 7, Figure 9, and Figure 10), there is sufficient reason to believe that a follow-on investigation may be able to identify a method for classification that is most appropriate for $P_D = 1$.

Literature Cited

- [1] Department Of Defense, "Unexploded ordnance response: Technology and cost", report to congress," March 2001.
- [2] L. Carin, N. Geng, M. McClure, Y. Dong, Z. Liu, J. He, J. Sichina, M. Ressler, L. Nguyen and A. Sullivan, "Wide-area detection of land mines and unexploded ordnance," *Inverse Problems*, vol. 18, pp. 575-609, 2002.
- [3] Y. Dong, P. R. Runkle, L. Carin, D. R., S. A., R. M. A. and S. J., "Multi-aspect detection of surface and shallow-buried unexploded ordnance via ultra-wideband synthetic aperture radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1259-1270, 2001.
- [4] L. M. Collins, Y. Zhang, J. Li, H. Wang, L. Carin, S. J. Hart, S. L. Rose-Pehrsson, H. H. Nelson and J. R. McDonald, "A Comparison of the Performance of Statistical and Fuzzy Algorithms for Unexploded Ordnance Detection," *IEEE Transactions on Fuzzy Systems*, vol. 9, pp. 17-30, 2001.
- [5] S. L. Tatum and L. M. Collins, "A Comparison of Algorithms for Subsurface Target Detection and Identification Using Time-Domain Electromagnetic Induction Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1299-1306, 2001.
- [6] S. Hart, H. Nelson, R. Grimm, S. Rose-Pehrsson and J. McDonald, "Probabilistic neural networks for unexploded ordnance (UXO) classification using data fusion of magnetometry and EM physics-derived parameters," in *UXO/Countermining Forum*, Anaheim, CA, 2000, .
- [7] B. Barrow and H. H. Nelson, "Model-Based Characterization of Electromagnetic Induction Signatures Obtained with the MTADS Electromagnetic Array," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1279-1285, 2001.
- [8] L. Carin, N. Dasgupta and H. Li, "Optimal sensor management for next-generation EMI systems," SERDP Project MM-1591 Final Report, 2008.

- [9] F. Morrison, T. Smith, A. Becker and E. Gasperikova, "DETECTION AND CLASSIFICATION OF BURIED METALLIC OBJECTS," SERDP Project UX-1225 Final Report, 2005.
- [10] K. O'Neill, "Processing for clutter evasion in UXO discrimination," SERDP Project MM-1590 Final Report, 2008.
- [11] F. Shubitidze, "Non-traditional physics-based inverse approaches for determining a buried Object's location," SERDP Project MM-1592 Final Report, 2008.
- [12] AETC, "PROCESSING TECHNIQUES FOR DISCRIMINATION BETWEEN BURIED UXO AND CLUTTER USING MULTISENSOR ARRAY DATA," SERDP Project CU-1121 Final Report, 2002.
- [13] J. Foley, Shaw Environmental and Sky Research, "Sensor orientation effects on UXO geophysical target discrimination," SERDP Project MM-1310 Final Report, 2006.
- [14] S. L. Tantom, Y. Yu and L. M. Collins, "Bayesian Mitigation of Sensor Position Errors to Improve Unexploded Ordnance Detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, pp. 103-107, 2008.
- [15] A. B. Tarokh and E. L. Miller, "Subsurface Sensing Under Sensor Positional Uncertainty," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 675-688, 2007.
- [16] F. Shubitidze, K. O'Neill, B. E. Barrowes, I. Shamatava, J. P. Fernandez, K. Sun and K. D. Paulsen, "Application of the normalized surface magnetic charge model to UXO discrimination in cases with overlapping signals," *Journal of Applied Geophysics*, vol. 61, pp. 292-303, 2007.
- [17] X. Chen, K. O'Neill, B. E. Barrowes, T. M. Grzegorzcyk and J. A. Kong, "Application of a spheroidal-mode approach and a differential evolution algorithm for inversion of magneto-quasistatic data in UXO discrimination," *Inverse Problems*, vol. 20, pp. S27-S40, 2007.
- [18] J. Stalnaker and E. Miller, "Particle swarm optimization as an inversion tool for a nonlinear UXO model," in *International Geoscience and Remote Sensing Symposium*, 2007, pp. 432.
- [19] L. Beran and D. W. Oldenburg, "Selecting a Discrimination Algorithm for Unexploded Ordnance Remediation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 2547-2557, 2008.
- [20] C. Elkan, "The foundations of cost-sensitive learning," in *International Joint Conference on Artificial Intelligence*, 2001, .
- [21] D. Pedro, "MetaCost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, United States, 1999, .

- [22] R. M. Bell and Y. Koren, "Lessons from the Netflix Prize Challenge," *SIGKDD Explorations*, vol. 9, pp. 75-79, 2007.
- [23] B. Harrison and P. Baggenstoss, "Hybrid discriminative/class-specific classifiers for narrowband signals," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, pp. 629-642, 2008.
- [24] S. Fine, J. Navratil and R. A. Gopinath, "Enhancing GMM scores using SVM "hints"," in *EUROSPEECH*, 2001, pp. 1757-1760.
- [25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [27] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148-156, 1996.
- [28] D. D. Snyder, D. C. George, S. C. MacInnes and R. L. Smith, "A COMPARATIVE ASSESSMENT OF SEVERAL DIPOLE-BASED ALGORITHMS FOR THE EXTRACTION OF UXO TARGET PARAMETERS," in *SAGEEP*, 2008, pp. 1390-1400.
- [29] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273-324, 1997.
- [30] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131-156, 1997.
- [31] R. N. McDonough and A. D. Whalen, *Detection of Signals in Noise*. San Diego: Academic Press, 1995.
- [32] J. J. Remus, K. D. Morton, P. A. Torriane, S. L. Tantom and L. M. Collins, "Comparison of a distance-based likelihood ratio test and k-nearest neighbor classification methods," in *IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 362-367.
- [33] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*. New York: Wiley, 2000.
- [34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [35] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.
- [36] A. Y. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Neural Information Processing Systems (NIPS)*, 2002, pp. 841-848.

Appendices

MATLAB Code: Model inversion outlier identification metric

```
function [OIM] = OutlierIdentificationMetric(Data,Features)
% [OIM] = OutlierIdentificationMetric(Data,Features)
%
% Calculates a correlation-based measure for identifying outliers from the
% model inversion process
%
% INPUTS:
% Data - a N x D matrix of data, where D is the number of dimensions (e.g.
% # of time samples multiplied by # of receivers) and N is the number of
% anomalies
%
% Features - a N x D1 matrix of features, where D1 is the dimensionality of
% the feature set and N is the number of anomalies as above
%
% OUTPUTS:
% OIM - a N x 1 column of outlier identification metrics, where higher
% values suggest a greater likelihood that the anomaly will be an outlier
% in feature space
%
% calculate correlation coefficients between individual measurements
rData = corrcoef(Data. ');

% calculate correlation coefficients between resulting features
% (normalized by standard deviation)
rFeats = corrcoef((Features./repmat(std(Features),size(Features,1),1)). ');

tempD = rData-rFeats;

% calculate percentage of anomalies for which correlation of data exceeds
% correlation of features
OIM = sum(tempD>0).'/size(tempD,1);
```


MATLAB Code: Distance likelihood ratio test (DLRT) classifier

```
function [LLRT,Yhat,PercentCorrect] = DLRTclassifier(Xtrain,Ytrain,Xtest,Ytest,K);
% [LLRT,Yhat,PercentCorrect] = DLRTclassifier(Xtrain,Ytrain,Xtest,Ytest,K);
%
% Implements the Distance Likelihood Ratio Test (DLRT) classifier.
%
% INPUTS:
% Xtrain - a D by N matrix of training data, where D is the number of
%          dimensions and N is the number of training samples
% Ytrain - a 1 by N vector of class labels {0,1}
% Xtest - a D by NN matrix of test data
% Ytest - a 1 by NN vector of class labels for the test data (if
%         available). If unavailable, input empty brackets []
% K - number of neighbors, must be a positive integer
%
% OUTPUTS:
% LLRT - the estimates of the likelihood ratio calculated by the DLRT
% Yhat - estimated class labels using the minimum-error threshold
% PercentCorrect - percentage of samples in Xtest correctly classifier (if
%                  "Ytest" was provided), scale 0 to 100
%
% Reference: Remus et al. "Comparison of a distance-based likelihood ratio
% test and k-nearest neighbor classification methods" Proceedings of the
% IEEE Workshop on Machine Learning in Signal Processing (MLSP) 2008
%
%
% initialize outputs
LLRT = zeros(1,size(Xtest,2));
Yhat = zeros(1,size(Xtest,2));

% find indices of the classes in the training data
classes = unique(Ytrain);
trainH1 = find(Ytrain==classes(2));
trainH0 = find(Ytrain==classes(1));

for i = 1:size(Xtest,2);
    distH1 = sort(sqrt(sum((Xtrain(:,trainH1) -
    repmat(Xtest(:,i),1,length(trainH1))).^2,1))); % distances from the i-th test point
    to all H1 training samples

    distH0 = sort(sqrt(sum((Xtrain(:,trainH0) -
    repmat(Xtest(:,i),1,length(trainH0))).^2,1))); % distances from the i-th test point
    to all H0 training samples

    LLRT(i) = log(numel(trainH0)/numel(trainH1)) +
    size(Xtrain,1)*log(distH0(K)/distH1(K)); % calculate the LLRT output
end

% estimate class labels using the minimum error threshold
Yhat(LLRT>=0) = classes(2);
Yhat(LLRT<0) = classes(1);

% calculate percent correct if Ytest is provided
if numel(Ytest) == size(Xtest,2);
    PercentCorrect = 100*sum(Yhat==Ytest)/length(Ytest);
else
    PercentCorrect = nan;
end
```