Receiver-Operating Characteristic (ROC) Curves to Assess Advanced Detection and Classification Technology for Environmental Remediation of Unexploded Ordnance (UXO)

Shelley Cazares[1], Jacob Bartel[1]
[1]Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311

Abstract
Receiver-operating characteristic (ROC) curves are often used to assess the performance of binary classification systems, allowing stakeholders to understand the trade-off between Type I (false positive) and Type II (false negative) errors. This works well in textbook cases. In real-world experiments, however, ROC curves can have unexpected subtleties that make them difficult to construct and interpret. For example, the Department of Defense is sponsoring the development of advanced platforms, sensors, algorithms, and processes to detect and classify unexploded ordnance (UXO) in the midst of clutter. UXO are duds—munitions that were previously armed and fired but failed to explode. UXO can still pose a risk of detonation even decades later, threatening the safety of nearby humans, animals, vegetation, and structures. Conducting blind tests to demonstrate detecting and classifying UXO is fraught with safety, logistical, and cost constraints that make it difficult to construct the textbook ROC curves. Yet with careful planning and a few key assumptions, ROC curves can still be crafted to quickly tell the story of how well advanced systems can detect and classify UXO versus clutter in terrestrial and underwater experiments.

Key Words: Environmental Remediation, Unexploded Ordnance (UXO), Receiver-Operating Characteristic Curve (ROC Curve), Probability of Detection (Pd), Probability of False Alarm (Pfa), False Alarm Rate (FAR)

1. Receiver-Operating Characteristic Curves

Receiver-operating characteristic (ROC) curves like Figure 1 are used to assess the performance of binary classification systems (Swets 1989). In the U.S. Department of Defense (DoD), this is often a matter of detecting and classifying *threats* among *clutter*. ROC curves provide a quick and intuitive way for stakeholders to understand the trade-off between Type I and II errors, otherwise known as false positives and false negatives. In many situations, this can be recast into a trade-off of:

- *Safety*, on the vertical axis of the ROC curve, using a metric like the probability of detection (Pd) to summarize how many threats are found, and
- *Unnecessary costs*, on the horizontal axis, using a metric like the probability of false alarms (Pfa) to summarize how many alarms are false.

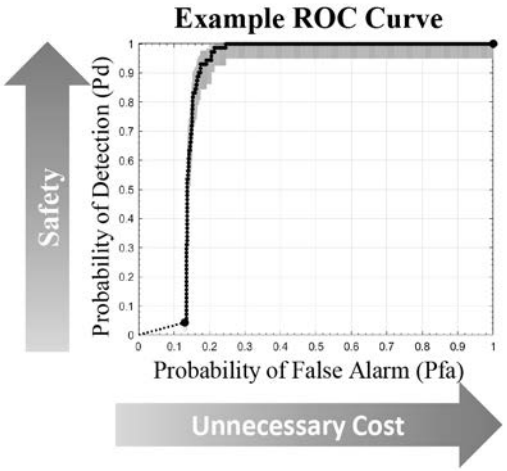

Figure 1. A receiver-operating characteristic (ROC) curve can summarize the trade-off between safety and unnecessary costs when detecting and classifying threats among clutter

This approach works well in textbook cases. In real-world experiments, however, ROC curves can have unexpected subtleties that make them difficult to construct and interpret. In particular, constructing ROC curves requires *full ground truth*—knowledge about which threats and clutter were truly present. Unfortunately, full ground truth can be difficult to obtain in some uses cases, such as the remediation of unexploded ordnance (UXO), especially in underwater environments.

In section 2, we first define UXO, explain how UXO can pose a threat to public safety, and describe the DoD's efforts to develop advanced systems to detect and classify UXO for safe removal or continued monitoring. In section 3, we explain how we have constructed ROC curves to demonstrate the performance of these advanced systems in remediating (cleaning up) *land* contaminated with UXO. In section 4, we explain why ROC curves are much more difficult to construct in *underwater* demonstrations and offer mitigations and assumptions needed to overcome these challenges. Finally, in section 5, we reflect on the need for rigorous procedures for collecting and ground-truthing data to enable further statistical analyses with ROC curves to support technology transition.

2. Faster, Safer Unexploded Ordnance Remediation

Thousands of formerly used defense sites and active installations in the United States are contaminated with UXO (USACE 2020). UXO are duds—munitions such as mortars, artillery, bombs, etc. that were previously armed and fired but failed to explode. Unfortunately, UXO can still pose the risk of detonation, even decades later, threatening the safety of nearby humans, animals, vegetation, and structures. Figure 2 shows photographs of UXO (left) and metallic clutter objects found near UXO on land (right).

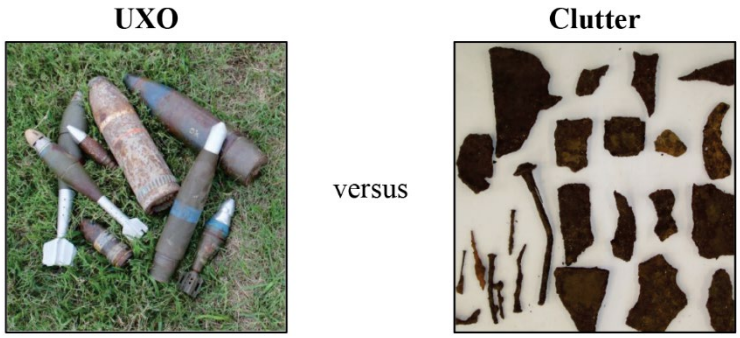

Figure 2: UXO (left) and metallic clutter objects (right) found near UXO on land. Reproduced from Cazares et al. (2021).

UXO can contaminate both land and underwater sites. Munitions can burrow into the ground, undetonated, and then later become uncovered during construction or farming efforts. Munitions can also be fired into oceans, rivers, and lakes, sinking through the water and reaching the seabed floor, undetonated; such underwater UXO can later wash onto the beach.

The risk of encountering UXO is often reported in the popular news. In 2008, the BBC reported on a World War II (WWII) era bomb found just outside London, United Kingdom. This UXO even "started to tick" (BBC 2008). Just a few years later, another WWII-era bomb was discovered in Koblenz, Germany. Half the town—45,000 residents—had to be evacuated so that the bomb could be safely disarmed (Day 2011). UXO is not simply a problem from the world wars, however. More and more UXO is being created each day. In 2012, CNN reported on UXO that the United States has left behind in Afghanistan, calling it "a dangerous legacy of war" (Jamjoom and Formanek 2012). Many areas in the United States are also contaminated with UXO, due to their prior uses as military training and test sites. For example, in 2014, UXO was found near Baltimore, Maryland and a 300-meter area had to be evacuated (Brown 2014). Just one month later, more UXO was found in the same metropolitan area, prompting an evacuation of another neighborhood (Wells 2014). UXO in underwater locations can also cause significant concern. In 2020, Polish military forces remotely neutralized a six-ton "Tallboy" that had been dropped by the British into a shipping canal off the Polish port city of Swinoujscie in WWII. This UXO had to be remotely detonated, "sending a plume of water high into the air" (Ismay 2020). Underwater UXO remediation can be so difficult that authorities sometimes choose to simply leave it in place, as has been recently debated in Hawaii (AP 2020).

The U.S. DoD seeks to remediate the threat of UXO. For over two decades, the DoD's Strategic Environmental Research and Development Program (SERDP) and Environmental Security Technology Certification Program (ESTCP) have funded the development of advanced platforms, sensors, algorithms, and processes for detecting and classifying UXO among clutter (Andrews and Nelson 2011; SERDP and ESTCP n.d.)

The DoD has also funded demonstrations (experimental tests) to assess the performance of these advanced systems. Between 2007 and 2017, ESTCP sponsored over 20 live-site demonstrations across the United States and territories in which the advanced UXO detection and classification systems were used on real terrestrial (land) sites contaminated with real munitions and explosives of concern, such as UXO (Andrews and Nelson 2011; Cazares, Ayers, and Tuley 2018). These live-site demonstrations were key to obtaining regulatory approval for the advanced systems and transferring them to commercial use.

The DoD then turned its focus to the underwater environment. In recent years, SERDP and ESTCP have funded the development of advanced platforms, sensors, algorithms, and processes for the underwater environment and have sponsored underwater technology demonstrations (Richardson and Bradley 2021). Designing, executing, and scoring these demonstrations is much more difficult underwater than on land, due to the additional safety, logistical, and engineering constraints of the underwater environment.

3. Terrestrial Demonstrations

Over the past two decades, the DoD has invested in the research and development of advanced systems to detect metallic objects buried in the ground and correctly classify them as UXO or clutter. For example, the left side of Figure 3 shows a photograph of the Time-domain Electromagnetic Multi-Sensor Towed Array Detection System (TEMTADS), developed by the Naval Research Laboratory and Nova Research. The TEMTADS uses electromagnetic induction (EMI) sensors to collect data over the ground as the operator pushes it in a lawnmower-like pattern, resulting in full coverage of the test site. The TEMTADS operators use physics-based signal-processing methods to detect buried metallic objects in the collected data and then extract features from the collected data around each detected object. Straightforward library-matching techniques are then often used to classify the features extracted for each detected object as "likely UXO" or "likely clutter" (Steinhurst et al. 2013).

Time-domain Electromagnetic Multi-sensor Towed Array Detection System (TEMTADS), developed by the Naval Research Laboratory and Nova Research

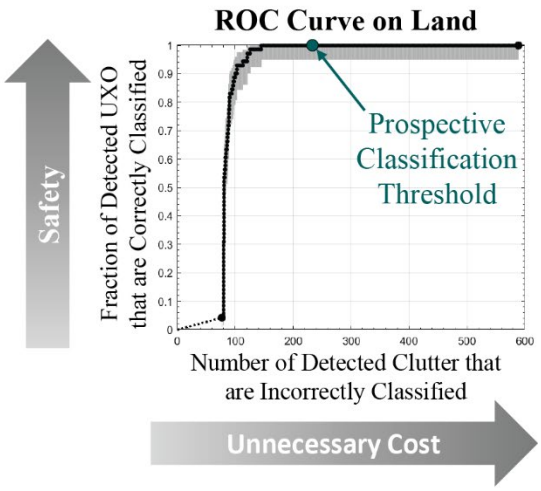

Figure 3: The Time-domain Electromagnetic Multi-sensor Towed Array Detection System (TEMTADS), developed by the Naval Research Laboratory and Nova Research to detect and classify UXO and clutter on land. Reproduced from Cazares et al. (2021).

The Institute for Defense Analyses (IDA) assisted in the design and scoring of experiments to demonstrate how well different systems can detect and classify UXO on land (Cazares et al. 2021; Fisher and Cazares 2021; Cazares, Ayers, and Tuley 2018). A key step involved excavation teams digging up and measuring every single object detected in the test site to collect ground truth. Ground truth is information about the true locations of all detected objects, along with their true labels—UXO or clutter. This ground-truthing process allowed ESTCP to confirm the locations and nature of *all* objects detected at the test site, including

the UXO emplaced for the purpose of the demonstration as well as the UXO present at the test site before we even arrived—remnants of the area's previous use as a military training or test site. Of course, this process did not reveal ground truth about any objects that were *not* detected. However, since multiple systems were used at each terrestrial demonstration—some systems even employed by multiple demonstration teams—it is likely that all UXO was detected by at least one system employed by at least one team.

There were two types of terrestrial demonstrations. Early demonstrations in the late 1990s and early 2000s were conducted on carefully controlled *standardized sites* that were fully cleared of all metallic objects (both UXO and clutter) before emplacing dozens of inert and surrogate UXO, as well as dozens of clutter objects, for the demonstration. In contrast, later terrestrial demonstrations, such as those described here from 2007 to 2017, were conducted on *live sites*—land potentially contaminated with real UXO due to previous use as military training and test sites (Cazares et al. 2021; Fisher and Cazares 2021; Cazares, Ayers, and Tuley 2018). To increase sample size, dozens of additional inert or surrogate UXO were emplaced at the live sites, using the same types of UXO that were originally fired at the sites according to historical military records. This increased sample size gave the systems more opportunity to detect and classify UXO in the demonstration. The live sites already contained hundreds of metallic clutter objects that had been dropped or discarded over the years before we even arrived to set up the demonstration, such as old tools, broken plow pieces, balls of barbed wire, fragments of previously exploded munitions, etc. Therefore, additional clutter objects were *not* emplaced at the live sites since their sample size was already high.

To score the demonstrations, IDA compared the ground truth with each system's alarms to construct ROC curves, such as that shown on the right side of Figure 3—one of the many ROC curves generated to assess the performance of the TEMTADS on land. Each point on the ROC curve corresponds to a different classification threshold that the system could have used to differentiate between "likely UXO" and "likely clutter" objects in this particular demonstration. A vertical gray line is drawn through each point to indicate the 95% confidence interval around that point's Pd value. In this ROC curve, the individual gray confidence intervals are so numerous that they overlap at this scale to appear as a band. Each individual confidence interval was calculated with the Clopper-Pearson method, using the beta distribution as the standard conjugate prior for the binomial distribution (Brown, Cai, and DasGupta 2001), with no adjustment for multiple comparisons. The blue dot is the prospective classification threshold—the threshold selected by the system developers during the experiment, with no knowledge of ground truth.

The blue dot on this ROC curve can quickly tell an informative story about the system performance: This system was able to correctly classify 100% of the UXO (i.e., at the top of the vertical axis) while generating only 234 false alarms, less than half (i.e., less than halfway to the right along the horizontal axis) of the almost 600 false alarms that would have otherwise been produced if no classification algorithms were used at all.

Between 2007 and 2017, IDA constructed several hundred ROC curves like these to summarize the results of over a dozen systems at 20 live-site demonstrations—real sites potentially contaminated with real UXO (Cazares et al. 2021, Cazares, Ayers, and Tuley 2018). Later live-site demonstrations, which were intentionally designed to be increasingly challenging, typically included smaller, more varied UXO emplaced closer to clutter objects, as well as more difficult terrain (sloped surfaces and thick vegetation to challenge

the platforms' mobility, thick tree cover to challenge the geolocation equipment, and soil with high iron content that led to high noise floors for the EMI-based sensors).

Despite these challenges, by 2012, most demonstrators at many live sites showed good to excellent performance, detecting and correctly classifying 100% of the UXO while incorrectly classifying fewer than 30% of the detected clutter objects (Cazares et al. 2021). At some sites, however, results were not as good. These varied results allowed the DoD to better understand which types of sites were better suited for the use of this technology. Simple, straightforward ROC curves, like that shown in Figure 3, could quickly summarize a system's performance in particular environments, easing communication among environmental regulators, remediation teams, landowners, and the general public about the opportunities and limitations of advanced UXO detection and classification systems on land.

4. Underwater Demonstrations

The DoD's research teams have turned their eyes to the sea. Advanced systems are now being developed to deploy sensors in the water, such as the Multi-Sensor Towbody (MuST). The MuST, circled in yellow on the left side of Figure 4, has been developed by the Applied Physics Laboratory of the University of Washington. The MuST uses synthetic aperture and sidescan sonar to collect data while being towed by a boat through the water in an overlapping pattern to achieve full coverage of the seabed floor, often at multiple azimuthal angles (Williams et al. 2021). IDA researchers are now assisting with the design and scoring of experiments to demonstrate how well underwater systems like the MuST can detect and classify UXO that are lying on or buried in the seabed floor.

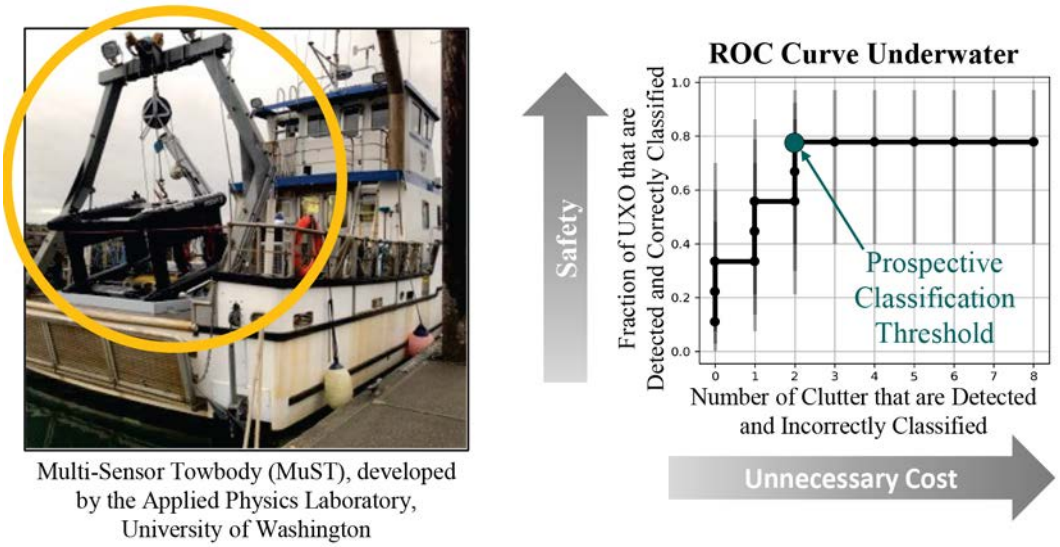

Figure 4: The Multi-Sensor Towbody (MuST), developed by the Applied Physics Laboratory of the University of Washington, to detect and classify UXO versus clutter underwater. Reproduced from Williams et al. (2021).

Unfortunately, designing, executing, and scoring these experiments is much more difficult underwater than on land, due to the additional safety, logistical, and engineering constraints of the underwater environment. In short, ROC curves are much more difficult to construct for underwater demonstrations. The underwater challenges lead to schedule and budget

constraints that limit the number of objects that can be emplaced at the test site. The reduced sample size leads to ROC curves that are less smooth and have wider confidence intervals, as can be seen by comparing the Figure 3 ROC curve (from a terrestrial live-site demonstration with over 70 detected UXO objects and over 600 detected clutter objects) with the Figure 4 ROC curve (from an underwater standardized-site demonstration with only nine emplaced UXO objects and six emplaced clutter objects). Other, more subtle difficulties also emerge in the underwater environment. Many of the lessons learned from the terrestrial demonstrations can be leveraged to address some of the challenges of the underwater environment. In some cases, though, new solutions have been found to overcome the added challenges of designing, executing, and scoring underwater demonstrations. Table 1 lists each of the main challenges, along with the mitigations and assumptions that have been used so far to overcome them.

Table 1: Challenges in Underwater UXO Detection and Classification Demonstrations and the Mitigations and Assumptions to Address Them

	Challenge	***Mitigation/Assumption***
1	Greater regulatory hurdles (permit restrictions due to shipping lanes, fishing and crabbing seasons, and flora/fauna protections)	• Carefully select the test site and season • Reduce the time objects are emplaced in the seabed floor • Plan for long lead times to acquire permits
2	Potential movement of emplaced objects between ground truthing and data collection (due to waves and currents, dragging by ship anchors, etc.)	• Carefully select the test season • Reduce the time between ground truthing and data collection • Assume objects do not move in the time between ground truthing and data collection • Collect ground truth before and after data collection • Instrument "smart" objects to track their movement over time
3	Less precise geolocation instruments (for ground truth and system alarms)	• Larger detection halo radius R must be used • All UXO must be emplaced ≥2R apart from each other
4	More limited resources (time, funds) to excavate detected objects that were *not* emplaced	• Assume the only UXO objects in the test area were those that were emplaced • Assume the only clutter objects in the test area were those that were detected but *not* emplaced UXO • Plot number of false alarms or false-alarm rate (FAR) instead of Pfa on the horizontal axis of the ROC curve
5	Diver safety concerns during emplacement and ground truthing	

3.1 Greater Regulatory Hurdles

There are often much greater regulatory hurdles when operating underwater, due to permit restrictions to avoid shipping lanes and fishing or crabbing seasons, as well as to protect underwater flora and fauna, including coral, kelp beds, fish, and other marine life. There are a few ways to overcome these hurdles.

- First, the test site must be carefully selected, as well as the month or season of the year in which the demonstration is conducted, to avoid times and locations that are highly regulated.
- Once the test site and season are selected, the first step in any demonstration is to have a scuba-diving team emplace objects in or on the seabed floor, so that the advanced systems can have the opportunity to detect and classify them. These

emplaced objects include UXO (for safety, only inert and surrogate UXO are used in demonstrations) and clutter objects (old scuba tanks, ship anchors, crab pots, etc.) Permit restrictions may limit the length of time those objects can be left in or on the seabed floor. Therefore, the more the demonstration schedule can be compressed, the better.

- Long lead times to acquire the necessary permits should be built into the demonstration planning schedule.

Of course, these mitigations can lead to increased demonstration costs. Much more funding is needed to conduct underwater demonstrations than on land, and these mitigations are just a few reasons why.

3.2 Potential Movement of Emplaced Objects between Ground Truthing and Data Collection

The objects emplaced in the test site for the demonstration can lead to other challenges, as well. On land, the emplaced objects stay put. Underwater, though, they can move, particularly in energetic environments like the surf zone, where there are breaking waves and strong currents. The objects' movement could occur anytime between their emplacement and removal. Of particular concern, though, are cases in which objects move in between collecting ground truth and collecting the data used for detection and classification. If the objects move in between ground truthing and data collection, then the ground truth is no longer "true," and it becomes very difficult, if not impossible, to objectively score the demonstration.

There are different ways to mitigate this risk:

- We can carefully choose the month or season of the demonstration to make it less likely that the wave action or water currents will be strong enough to uncover the objects and move them over the sediment. Physics-based models for underwater UXO burial, reemergence, and migration due to wave action or water currents can be particularly helpful in selecting the appropriate season for the demonstration (Klammler, Sheremet, and Calantoni 2020; Bruder, Cristaudo, and Puleo 2020). Selecting the ideal season does introduce a bias into the results of the demonstration, since the results of a demonstration conducted at an ideal time of the year cannot necessarily extrapolate to other times of the year. However, this bias is not necessarily problematic. The operational tempo of UXO remediation projects is much less demanding than in wartime military counter-explosive missions. UXO often rest in place for years or decades, and so remediation teams can afford to wait the needed weeks or months to select the ideal conditions for UXO remediation.
- We can also compress the demonstration schedule, leaving little opportunity for the emplaced objects to move in between the ground truthing and the data collection for detection and classification.

These two mitigations above are similar to those already employed to address the additional regulatory hurdles of operating in the underwater environment. Even with these mitigations, though, there is still a residual risk that the emplaced objects can move. To press forward with scoring, then, we must do at least one of the following:

- Assume that the emplaced objects do *not* move in the time between ground truthing and data collection. However, this assumption may not always be suitable for energetic environments like the surf zone.

- Collect ground truth both before *and* after data collection. Although this would require more time and funding, any discrepancies between the two ground truth sets could indicate that an object has moved. However, we would not know *when* the object moved (before, after, or even during the data collection), and therefore we would not know which ground truth set, if any, would be appropriate for use in scoring. Instead, we would have to remove those objects from further analysis, which would be unfortunate since sample size is already limited.
- Instrument the emplaced objects so that we can track their movements over time (Frank, Landry, and Calantoni 2016; Bruder, Cristaudo, and Puleo 2020). This is easier to do with larger objects (e.g., 105 mm versus 37 mm projectiles), since more space is available inside larger (inert) munitions for the required electronics.

3.3 Less Precise Geolocation Instruments

The next challenge is technical in nature. Less precise geolocation equipment is available for deeper underwater environments. In terrestrial demonstrations, a real-time kinematic (RTK) Global Positioning System (GPS) can be used to collect extremely precise information about the locations in which the objects were emplaced (the ground truth) and the locations detected by the system during the demonstration (the alarms). However, like all radio frequency transmissions, GPS signals do not travel well through water (Taraldsen, Reinen, and Berg 2011). Therefore, different types of geolocation instruments must be used in deeper underwater demonstrations, and these instruments usually have precision levels that can be at least one order of magnitude worse than on land (Woodruff et al. 2021). Although it can be possible to emplace large, carefully surveyed monuments to aid in geolocation, the permitting restrictions discussed above do not always allow for their use, and they do not eliminate the problem entirely.

As a result, scoring schemes for underwater UXO detection and classification demonstrations must be flexible enough to handle this increased geolocation error. Figure 5 shows a bird's-eye view of a notional test site. In this cartoon example, an underwater system collected data over the test area during the demonstration, processed those data, and generated four alarms (×)—four detected objects classified as "likely UXO."

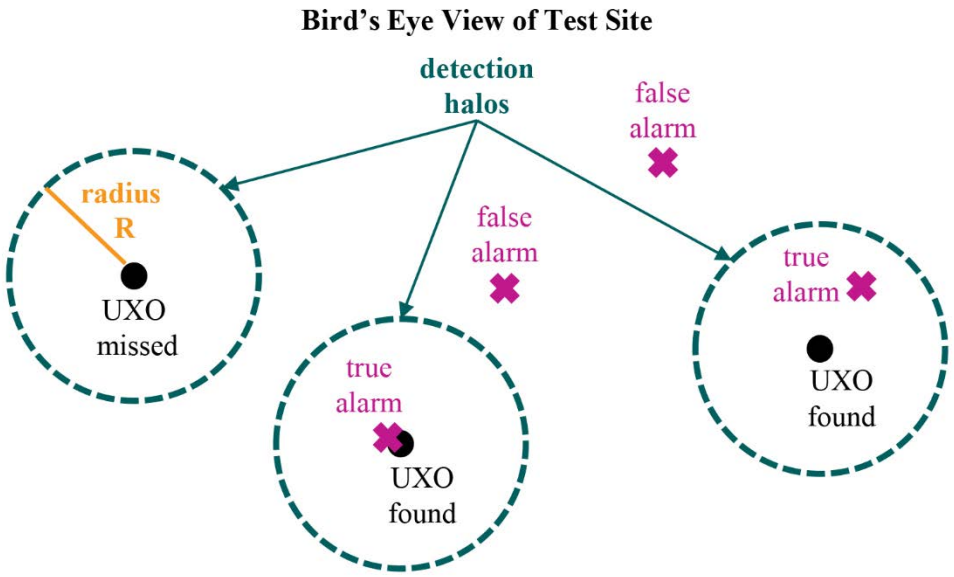

Figure 5: Detection halos (---) of radius R used for scoring UXO detection and classification demonstrations. UXO (•) is scored as "found" (a true positive) if there is at least one alarm (×) within a distance R, otherwise as "missed" (a false negative). False alarms (false positives) are those alarms (×) farther than a distance R from any UXO (•).

The performance of this notional system can be scored by using detection halos (---) to compare the system alarms (×) to ground truth (•)—the locations in which inert UXO were truly emplaced. Detection halos are imaginary circles of radius R centered on each emplaced UXO object (•). We score UXO (•) as "found" (a true positive) if an alarm (×) is within a distance R, such as the two true alarms matching up with the two found UXO objects in the middle and right parts of Figure 5. In this simple cartoon, we also have two false positives—the two false alarms in the top part of the figure. In addition, we have one false negative in the left part of the figure: One out of the three UXO objects was missed, resulting in a Pd of 2/3, or 67%.

Selecting R, the radius of the detection halos, can be one of the trickiest steps in designing and scoring an underwater UXO remediation demonstration. In terrestrial demonstrations, R could theoretically be as small as a few centimeters. In practice, though, R is often set to be 20–30 cm on land, approximately the width of the remediation team's shovel when carefully excavating a UXO for safe removal (Cazares, Ayers, and Tuley 2018). In underwater demonstrations, though, R must be much larger, due to the increased geolocation error.

Furthermore, to avoid complex corner cases in scoring (e.g., double-counting true or false positives), it can also be helpful to ensure that all UXO are emplaced farther than 2R away from each other. As technology matures, and underwater systems become able to detect and classify more closely spaced UXO, this constraint will need to be relaxed. Therefore, future demonstrations will require additional scoring rules to handle these corner cases.

3.4 More Limited Resources and Diver Safety Concerns

The last two challenges are interrelated. In a terrestrial demonstration, it is possible to obtain the necessary time and funding to excavate every single object detected in the test area to construct full ground truth. In an underwater demonstration, however, it is largely

impossible to amass the necessary funding and block out the required schedule to do so. Furthermore, there are safety concerns regarding how much time the scuba divers can spend underwater to perform this function. As a result, full ground truth is unlikely to be available for underwater demonstrations, and we therefore must reconsider our definition of the metrics we plot on the vertical and horizontal axes of our ROC curves.

Traditionally, Pd is plotted on the vertical axis of the ROC curve. (Pd is called "sensitivity" in the medical literature and "recall" in the information-retrieval literature.) As shown in Equation 1, Pd is defined as the probability that the system will deliver an alarm when a true UXO object is present. Using the terminology from the notional test site in Figure 5, this is often calculated as the number of true alarms divided by the number of instances in which a true UXO was present, that is, the number of true UXO (Swets 1989). However, it is impossible to know how many UXO are truly present at a test site. Of course, we know how many UXO were emplaced at the site for the purpose of the demonstration. Unfortunately, we do not know if there were any other UXO present at the site before we even arrived—remnants of previous military training and test sites. Although historical records may provide some information about which types of munitions (if any) were previously fired at a military site, there are many cases in which the records are missing or incomplete. Furthermore, almost no records contain the level of detail needed to pinpoint the individual UXO objects. Yet by carefully selecting the location of our test sites, we can make good assumptions about the denominator of our Pd metric—namely, that the only UXO present at the test site were those that were emplaced for the demonstration.

$$\text{Pd} = \text{p}(\text{Alarm} \mid \text{True UXO}) = \frac{\#\ \text{True Alarms}}{\#\ \text{True UXO}} \approx \frac{\#\ \text{True Alarms}}{\#\ \text{Emplaced UXO}}$$

Equation 1

Such assumptions are still possible to make, since SERDP and ESTCP are still at the early stages of using *standardized sites* for underwater UXO demonstrations—carefully controlled sites that are prepared for the purpose of demonstration. Standardized sites can be specifically selected for having *not* been previously used as military training and test sites, making it unlikely that any UXO was already present in the test area before we even arrived for the demonstration. However, as underwater technology improves, more challenging demonstrations will be needed, such as *live-site* demonstrations in underwater areas potentially contaminated with real UXO. For future underwater live-site demonstrations, the assumption inherent in the denominator of Equation 1 may no longer be valid, and other mitigations may need to be found.

The horizontal axis of the ROC curve is even trickier for underwater demonstrations. Traditionally, Pfa is plotted on the horizontal axis of the ROC curve. (Pfa is called "1-specificity" in the medical literature.) Theoretically, Pfa is meant to be the probability that the system will generate an alarm when true UXO is *not* present. Using the Figure 5 terminology, this is often calculated as the number of false alarms divided by the number of instances in which UXO was *not* present (Swets 1989), as shown in Equations 2 and 3. In practice, though, Pfa can only be calculated for pure classification tests, such as the case in which one has already detected objects in the seabed floor and it is now simply a matter of classifying each of those detected objects as UXO or *not* UXO (such that one can count true negatives as well as false positives). However, it can be useful to define a metric to assess both the classification *and* detection capability of a system, since both tasks are

important to the overall goal of UXO remediation. In such cases, we can still define metrics that are similar in spirit to Pfa. Each has its own pros and cons, discussed below.

Equation 2 borrows a similar assumption from Equation 1. The denominator assumes that the only detected objects that were truly *not* UXO were clutter objects emplaced for the purpose of the demonstration. Unfortunately, that is often a poor assumption. Many test sites have a large number of naturally occurring clutter objects, including rocks, trash, discarded anchors, and fragments of previously exploded munitions.

$$\text{Pfa} = \text{p}(\text{Alarm} \mid \text{No True UXO}) = \frac{\#\text{ False Alarms}}{\#\text{ No True UXO}} \approx \frac{\#\text{ False Alarms}}{\#\text{ Emplaced Clutter}}$$

Equation 2

Equation 3 employs a different assumption to estimate the number of detected objects that were truly *not* UXO. The denominator of Pfa is now the total number of detected objects, minus the total number of emplaced UXO. That is, Equation 3 assumes that the only clutter objects in the test area were those that were detected but were *not* emplaced UXO. This is often a fair assumption to make for standardized sites, and so Equation 3 has often been used as a Pfa metric for assessing the capability of advanced systems for UXO detection and classification in underwater environments.

$$\text{Pfa} = \text{p}(\text{Alarm} \mid \text{No True UXO}) = \frac{\#\text{ False Alarms}}{\#\text{ No True UXO}} \approx \frac{\#\text{ False Alarms}}{\#\text{ Detections} - \#\text{ Emplaced UXO}}$$

Equation 3

However, the denominator's estimate in Equation 3 is inherently biased by the system's detection threshold:

- A detection algorithm with a low threshold will detect more objects (leading to a large denominator). Coupled with a good classification algorithm, this system will result in a small number of false alarms (leading to a small numerator), resulting in an artificially small (good) Pfa (produced by the small numerator divided by the artificially large denominator).
- In contrast, a detection algorithm with a high threshold will detect fewer objects (leading to a small denominator). *Coupled with the exact same good classification algorithm as above,* this system will result in an artificially large (poor) Pfa (produced by the small numerator divided by the artificially small denominator).

As a result of this bias, the Pfa metric in Equation 3 cannot be used to compare two different systems based on two different detection algorithms. Instead, Equation 3 is best used to assess a single system, since it conveys how well the system's classification algorithm cleaned up the problematic detections made by its own detection algorithm—a relative measure of the value-added of the classification algorithm.

In contrast, Equation 4 provides an absolute measure of the system's value as a whole, including both its detection and classification capabilities. The denominator is dropped and the numerator is used by itself: the number of false alarms. This metric can be easily converted to the unnecessary cost of employing the system for UXO detection and classification—one can multiply the number of false alarms by the dollar cost of further investigating an alarm (i.e., the dollar cost of sending divers down to excavate and remove the object or the dollar cost of carefully monitoring the object to determine if it has started

to wash up onto the beach after a storm—only to later find that the object was never explosive in the first place). Equation 4 was used to define the horizontal axes of both the terrestrial and underwater ROC curves on the right sides of Figures 3 and 4.

$$\text{FA} = \#\ \text{False Alarms}$$

Equation 4

Finally, Equation 5 normalizes the number of false alarms by dividing by the physical size of the test site. The false alarm rate (FAR) conveys the number of false alarms per unit area of the test site. FAR allows comparison between different demonstrations of different sizes.

$$\text{FAR} = \frac{\#\ \text{False Alarms}}{\text{Test Area}}$$

Equation 5

With all of these mitigations and assumptions in place, ROC curves can still be crafted to quickly tell the story of how well advanced systems can detect and classify UXO versus clutter in underwater environments, such as that shown on the right side of Figure 4. Comparing the ROC curves in Figures 3 and 4 (terrestrial and underwater), we see that the larger vertical steps and the wider vertical confidence intervals in Figure 4 (underwater) reveal that fewer UXO were emplaced in the test site for the underwater demonstration, due to the many safety, logistical, and cost constraints described above. Nevertheless, the blue dot on the Figure 4 ROC curve tells us that the system could correctly detect and classify 78% of the emplaced UXO objects while generating only two false alarms, an imperfect but promising balance between safety and unnecessary costs. A failure analysis showed that two of the nine (22%) emplaced UXO objects were not detected; these were both 81 mm projectiles, the smallest UXO objects in the demonstration. Further research and development are ongoing.

5. Avoiding "Garbage In, Garbage Out" Statistics

ROC curves are useful statistical tools for assessing the trade-off between false positives and negatives in detection and classification problems. Unfortunately, constructing ROC curves requires full ground truth, which can be difficult to obtain for some use cases, such as underwater UXO remediation. Despite the many challenges to conducting underwater demonstrations, several mitigations and assumptions can be employed to allow for the construction and easy interpretation of ROC curves.

The DoD is now developing four underwater testbeds for early demonstration of underwater UXO detection and classification systems, as illustrated in Figure 6. A preliminary demonstration was conducted in late 2020 in Sequim Bay, WA (Williams et al. 2021), and two more are scheduled for late 2021 with larger sample sizes. Other demonstrations are anticipated in Moku o Lo'e, HI; Panama City, FL; and La Spezia, Italy (Richardson and Bradley 2021).

Figure 6: Four underwater testbeds to demonstrate advanced detection and classification systems for underwater UXO remediation

Each of these demonstrations will face the inevitable challenges of the underwater environment. Domain expertise of underwater operations will be needed to identify and employ the appropriate mitigations and assumptions required for obtaining ground truth. Careful ground truthing will allow for rigorous scoring, avoiding the "garbage in, garbage out" problem that could otherwise result in ill-formed or misinterpreted ROC curves. With the appropriate mitigations and assumptions in place, solid ROC curves can be constructed to support painstaking failure analyses, detailed discussions across the UXO remediation community, and spirited debates between system developers, operators, policymakers, and regulators. All of these activities will be needed to transition the novel research and development into trusted commercial use.

Acknowledgments

We thank David Bradley and Herbert Nelson at SERDP and ESTCP for their financial sponsorship of this work. We also thank Michael Richardson, Michael Tuley, Steven Rabinowitz, Leon Hirsch, and John Biddle at IDA for their careful review throughout this work.

References

Andrews, Anne, and Herb Nelson. 2011. *Implementing Advanced Classification on Munitions Response Sites: A Guide to Informed Decision Making for Project Managers, Regulators, and Contractors. Environmental Security Technology Certification Program*. Accessed September 18, 2021, https://serdp-estcp.org/content/download/12780/151578/file/Implementing_Classification_on_Munitions_Response_Sites_FR%20with%20Appendix%20A.pdf.

Associated Press. 2020. Hawaii plans to leave Maui underwater war ordnance in place. *U.S. News*, August 21, 2020. Accessed September 18, 2020. https://www.usnews.com/news/best-states/hawaii/articles/2020-08-21/hawaii-plans-to-leave-maui-underwater-war-ordnance-in-place.

BBC. 2008. Unexploded bomb "started to tick," June 5, 2008. Accessed September 18, 2021. http://news.bbc.co.uk/2/hi/uk_news/england/london/7437086.stm.

Brown, Matthew Hay. 2014. Unexploded ordnance shuts down part of Fort Meade. *The Baltimore Sun*, March 11, 2014. Accessed September 18, 2021. http://articles.baltimoresun.com/2014-

03-11/news/bs-md-fort-meade-unexploded-ordnance-20140311_1_ordnance-utility-workers-300-meter-area.

Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta. 2001 Interval estimation for a binomial proportion. *Statistical Science* 16 (2): 101–33. doi:10.1214/ss/1009213286.

Bruder, Brittany, Demetra Cristaudo, and Jack A. Puleo. 2020. Smart surrogate munitions for nearshore unexploded ordnance mobility/burial studies. *IEEE Journal of Oceanic Engineering* 45(1). doi:10.1109/JOE.2018.2871227.

Cazares, Shelley, Elizabeth Ayers, and Michael Tuley. 2018. *ESTCP UXO Live Site Demonstrations 2007 to 2017*. IDA Document D-9193. Alexandria VA: Institute for Defense Analyses.

Cazares, Shelley, Elizabeth Ayers, Katherine Fisher, and Michael Tuley. 2021. Through the Valley of Death: The live site demonstrations for advanced geophysical classification of terrestrial unexploded ordnance. *FastTIMES* 26(1). Environmental and Engineering Geophysical Society. Accessed September 18, 2021, https://fasttimesonline.co/the-live-site-demonstrations-for-advanced-geophysical-classification-of-terrestrial-unexploded-ordenance/.

Day, Matthew. 2011. Second World War bomb leads to mass evacuation of German town. *The Daily Telegraph*, November 28, 2011. Accessed September 18, 2021. http://www.telegraph.co.uk/history/world-war-two/8920347/Second-World-War-bomb-leads-to-mass-evacuation-of-German-town.html.

Fisher, Katherine and Shelley Cazares. 2021. Assessing the capability of advanced geophysical classification (AGC) to inform minimum separation distance (MSD) in unexploded ordnance (UXO) remediation. *Symposium on the Application of Geophysics to Engineering and Environmental Problems (SAGEEP) 2021*, virtual, March 14–29, 2021. https://library.seg.org/doi/epdf/10.4133/sageep.33-109.

Frank, Donya, Blake J. Landry, and Joseph Calantoni. 2016. Investigating munitions mobility in oscillatory flows with inertial measurement units. *OCEANS 2016 MTS/IEEE*, Monterey CA, September 19–23, 2016, pp. 1–6. doi:10.1109/OCEANS.2016.7761158.

Ismay, John. 2020. World War II-era "earthquake bomb" exploded in Polish waters. *The New York Times*, October 14, 2020. Accessed September 18, 2020. https://www.nytimes.com/2020/10/14/world/europe/poland-bomb.html.

Jamjoom, Mohammed and Formanek, Ingrid. 2012. Unexploded munitions a dangerous legacy of war in Afghanistan. *CNN*, June 8 2021. Accessed September 18, 2021. http://www.cnn.com/2012/06/08/world/asia/afghanistan-unexploded-ordnances/index.html#.

Klammler, Harald, Alex Sheremet, and Joseph Calantoni. 2020. Seafloor burial of unexploded ordnance by wave-induced sediment instability. *IEEE Journal of Oceanic Engineering* 45(3). doi:10.1109/JOE.2019.2919356.

Richardson, Michael, and David Bradley. 2021. SERDP/ESTCP Munitions Response Program: An update on underwater remediation of unexploded ordnance (UXO). *Underwater Acoustics Conference and Exhibition (UACE) 2021*, virtual, June 21–24, 2021.

SERDP and ESTCP. No date. *Munitions response fact sheet*. Strategic Environmental Research and Development Program and Environmental Security Technology Certification Program. Accessed September 18, 2021. https://serdp-estcp.org/content/download/51899/510776/file/MR%20Fact%20Sheet.pdf.

Steinhurst, Daniel, Thomas Bell, Bruce Barrow, James Kingdon, and Glenn Harbaugh. 2013. UXO detection and classification from dynamic survey data with advanced sensors: The TEMTADS MP 2x2 cart. *Symposium on the Application of Geophysics to Engineering and Environmental Problems (SAGEEP) 2013*, Denver CO, March 17–23, 2013. https://library.seg.org/doi/10.4133/sageep2013-046.1.

Swets, John A. 1989. Measuring the accuracy of diagnostic systems. *Science* 240 (4857): 1285–93. doi:10.1126/science.3287615.

Taraldsen, Gunnar, Tore Arne Reinen, and Tone Berg. 2011. The underwater GPS problem. *OCEANS 2011 IEEE*, Santander Spain, June 6–9, 2011, pp. 1–8. doi:10.1109/Oceans-Spain.2011.6003649.

USACE. 2021. *Formerly used defense sites program fact sheet.* U.S. Army Corps of Engineers. Accessed September 18, 2021. https://www.usace.army.mil/Media/Fact-Sheets/Fact-Sheet-Article-View/Article/1910599/formerly-used-defense-sites-program-fact-sheet/.

Wells, Carrie. 2014. WWII-era unexploded ordnance found at Fort Meade. *The Baltimore Sun*, April 4, 2014. Accessed September 18, 2021. http://articles.baltimoresun.com/2014-04-04/news/bs-md-fort-meade-ordnance-20140404_1_fort-meade-workers-300-meter-area-ordnance.

Williams, Kevin, Timothy Marston, Dana Woodruff, and Shelley Cazares. 2021. Results from an informal demonstration of a buried-UXO detection, classification, geo-location system. *Underwater Acoustics Conference and Exhibition (UACE) 2021*, virtual, June 21–24, 2021.

Woodruff, D., J. Haxel, J. Vavrinec, S. Southard, S. Zimmerman, and K. Hall. 2021. *Sequim Bay underwater UXO prototype demonstration site: Field operations summary 2020.* Pacific Northwest National Laboratory. Accessed September 18, 2021. https://www.osti.gov/servlets/purl/1784318.