

Deep Mix: AI in Littoral Sonar Operations

Lauri Vasankari^{1,2}, Adrian Borzyszkowski², Luca Zelioli² and Jukka Heikkonen²

Received: 15 February 2025 / Accepted: 17 March 2025

© Harbin Engineering University and Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract

The mine countermeasure operations consist of hunting i.e. searching for possible mines from the seabed and water column as well as demolition or disposal of found mines to enable safe passage of other naval units. Mine hunting is a task conducted by specialized naval units aiming to clear sea areas from estimated or realized naval mine threats. The search is usually initiated by sweeping broad areas with a side scan sonar to produce an initial picture of the seabed and localize possible threats for closer examination. The manual analysis process is challenging and time-consuming, especially in environments where seabeds are littered with mine-like objects such as rocks and other roughness. This paper researches the possibility of enhancing littoral mine hunting with object detection using a deep convolutional neural network and vision transformer methods separately and as a Mixture of Experts system to do the preliminary object detection and classification of the data multiples faster than a sole sonar operator. The results indicate that even a processed sonar image can be used to enhance mine hunting operations in a transparent, explainable manner but would benefit from enriching with additional data and signal analysis to ensure the detection of the most difficult targets.

Keywords Marine sensors; Mine countermeasures; Sonar and transducers; Convolutional neural network; Transfer learning; Vision transformer; Mixture of experts

1 Introduction

This paper explores the use of Artificial Intelligence (AI), more specifically machine learning and deep learning, to enhance littoral naval mine countermeasure operations by applying novel deep learning solutions to underwater sonar imagery for classification and object detection. Initially, an overview of mine warfare and mine countermeasure operations is presented, followed by an explanation of the use of AI in naval mine warfare, the identified research gap and the research contribution, as well as a brief survey of earlier research. The following section 2 explains the methods, data and models in the context of the research problem, followed

by the display and analysis of the results in section 3.

1.1 Naval mine warfare and mine countermeasures (MCM)

As described by Melia (1991), mine warfare has been a relevant and impactful naval warfare branch for centuries. It is recognized as a devious threat to even the largest naval forces, effectively rendering otherwise supreme naval forces unable to conduct operations if MCM operations have not been able to counter the threat. In essence, mine warfare functions as a force multiplier as it can, at least, delay an opposing navy from acting. Schwarz (2014) from U.S. Naval War College describes naval mine threat as consisting of various types of mines, with different sensors and payloads. In general, mines are categorized into bottom mines that are lying in the seabed and moored mines that are floating in the water column while anchored into the seabed. Moored mines are most often contact mines, i.e., the ship has to make direct contact with the mine to detonate it, whereas bottom mines rely on sensors that detect the influence of magnetic, acoustic and pressure interference. The third option is to remotely detonate the mine.

The purpose of mine warfare is to hinder the opponent's ability to operate freely; the kinetic impact is often times a secondary object, as the mere threat of mines has the desired effect to a certain extent. This leads to the necessity

Article Highlights

- AI can enhance Baltic Sea Mine Countermeasure operations.
- Operationally relevant data can be leveraged to an extent with transfer learning.
- Mixture of Experts architecture shows promising results.
- Explainable AI support for operators is feasible.

✉ Lauri Vasankari
lauri.vasankari@gmail.com

¹ Department of Military Technology, National Defence University, Kadettikouluntie 7, Helsinki 00860, Uusimaa, Finland

² Department of Computing, University of Turku, Vesilinnantie 3, Turku 20014, Varsinais-Suomi, Finland

of MCM operations.

MCM operations can be categorized into five categories regarding mission objectives. These categories are exploratory, reconnaissance, breakthrough, attrition and clearance operations (Hibbert, 2006). Exploratory operations are used to determine whether mines are present in the area, while attrition is used to keep the threat of mine low by continuous operations. Reconnaissance is usually a follow-up of exploratory operations, which aims to define the limits of the mine threat area, while breakthrough operations aim to quickly reopen the channel (Broughton, 1998).

The technical tasks of MCM units can be categorized into surveying and mapping, mine hunting, mine sweeping and mine disposal (Broughton, 1998; Hibbert, 2006), corresponding to the aforementioned mission objectives. Mapping and surveying denote scanning of the seabed to create an overall impression of the environment. Mine hunting is the process of scanning an area of interest to either declare it free of mine threat or to produce detections for further examination and, e.g., disposal. This research framework is within the exploratory operations and aims to enhance the technical tasks of mapping, surveying and mine hunting.

MCM operations utilize a variety of sensors, but to scan large areas of seabed, Side Scan Sonar (SSS) (Flemming, 1976) and Synthetic Aperture Sonar (SAS) (Hayes and Gough, 2009) are currently most used by different navies. The sensor is usually towed behind the operating unit at some measurable depth in the water column, where two sets of transducers scan the seabed on both sides of the system. Figure 1 displays Flemming's visualization of the sound lobes emitted from the SSS (Flemming, 1976).

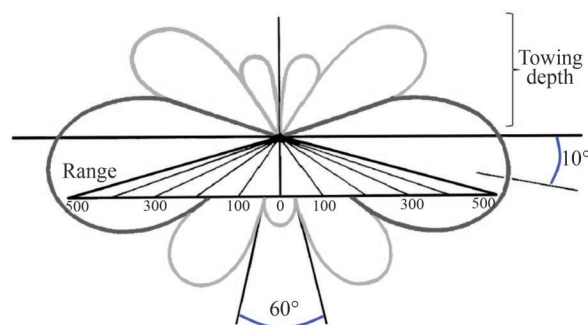


Figure 1 A schematic view of the sound lobes of an SSS (Flemming, 1976)

The mapping and survey are time-consuming, as the speed of the towing unit is limited when using an SSS (Flemming, 1976). High speed induces distortion in the resulting imagery and also affects the water depth of the towed array. Likewise, the area covered by one scan is limited to a few hundred meters per each side. While the data acquisition is time-consuming, so is the analysis of it to determine next actions. The operator has to manually process the imagery to detect objects of interest. A mapping or survey operation would produce a vast amount of imag-

ery to process, and hence automated object detection can greatly enhance the operator's ability to find the targets that need to be further examined.

The research of automation and AI in MCM operations has been driven by the stated time-consuming tasks, as well as the similarity of mine hunting pattern and object recognition tasks with existing computer vision and signal analysis solutions.

1.2 Research contribution

As Chapple (2008) has written, the purpose of automation and AI in mine hunting is to reduce the burden on human analysts in processing large volumes of sonar and other sensor data obtained by high-resolution sensors. Aided detection and classification can greatly speed up the MCM operations and free operators to other tasks such as disposal.

This work, using SSS data from the Baltic Sea region, collected by the Finnish Navy MCM units with Klein 5500 system (Armstrong et al., 1998), builds on earlier research presented in subsection 1.3. The dataset, described in detail in subsection 2.1, is novel and operationally relevant. The objects are classified as 'mine' if the information has been confirmed, whereas Mine-like Contact (MILCO) is an annotation that means that the operator perceives the object as a possible mine but it has not been confirmed with other methods. Other categories in this dataset include rocks and wrecks, see subsection 2.1 for examples. It is noticeable, however, that the data is not raw sensor data, as it is extracted from a system used for post-processing and reporting. Although the data set is limited in size, it is comparable, for example, with data sets obtained at the Coastal Systems Station in Panama City, Florida, which have been used by Hyland and Dobeck (1995), Chandran et al. (2002), and McKay et al. (2017). The oceanography of the Baltic Sea is a very challenging environment due to the uneven and rocky seabed (Kaskela, 2017), which is prone to induce a high number of false detections. This research extends earlier research by applying state-of-the-art Vision Transformer (ViT) (Dosovitskiy et al., 2021) to this detection and classification task, using a Support Vector Machine (SVM) (Boser et al., 1992) and deep convolution networks VGG16 and VGG19 (Simonyan and Zisserman, 2015) for comparisons that can be interlinked with earlier research. The primary contribution is testing a Mixture of Experts (MoE) (Masoudnia and Ebrahimpour, 2014) approach to utilize a selection of previously mentioned models in cohesion to further improve the classification and detection results.

As elaborated in subsection 1.3, the mine detection can be divided into two tasks, detection and classification. The research approach is to use the dataset for classification task and then leverage the classification models to aid in object detection. The used dataset does not directly support object detection in a verifiable manner, as the specific

object locations within images are not disclosed. Therefore, the only applicable approach is to use supervised learning to train an image classifier that can enable target localization through other methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). As explainability (Dwivedi et al., 2023) is an essential part of operational AI solutions, Grad-CAM and attention heat map provide visual representations of the deep learning model activation maps of identified class. The Grad-CAM uses gradients of a target concept to produce a coarse localization map of the region in the image used for the particular prediction, i.e., the output shows the final localization map and highlights the salient region of the prediction. The attention map of ViT (Brocki and Chung, 2023) works in a similar manner by using the attention scores of different regions in the image. Visualizing the perception and quality of the prediction introduces explainability to the used model. The resulting classifier models are evaluated as detection aids that are meant to support and enhance MCM operations.

The contribution in solving the research problem in this manner lies in demonstrating a low threshold deployment of research and development results: utilizing the same data currently used by the operators pinpoints the solution to fit into the existing *modus operandi*, promoting an easy to apply solution that could be quickly adapted into the operational theater. This research subsequently extends the use of Machine Learning (ML) methods by comparing convolution networks (LeCun and Bengio, 1998) to a vision transformer with attention mechanism (Vaswani et al., 2017) performance in the same detection and classification task. Furthermore, the use of MoE is proposed to take advantage of the different strengths of all the best performing models.

The main research contribution is producing an explainable classifier for mine hunting purposes in the littoral Baltic Sea waters using three applicable, state-of-the-art deep learning methods. Due to the recent, accumulating tensions in the area and increased military interest in the freedom of movement within the Baltic Sea (Hansen, 2025), the research topic is both relevant and impactful towards, e.g., NATO coalition forces conduct of operations.

1.3 Earlier research

The use of AI and ML been researched in MCM operation framework by, e.g., Mansell and Skinner (1995), since the 1980 s. Gorman and Sejnowski (1988) conducted early research on the topic in laboratory-like conditions aiming at binary classification of sonar echoes between a mine and a rock; the dataset is nowadays public but due to computational limitations of their time the data has been processed into a fixed number of channel features. Mansell and Skinner (1995) investigated decision-making support tools from propositional logic to neural networks and

Hyland and Dobeck (1995) have created Detection Density Adaptive Clutter Filter (ACF) for mine detection and classification from SSS images. Hyland and Dobeck (1995) claim 84% probability of detection and classification with a method that normalizes the image, applies ACF, selects the largest ACF pixels, convolves these into mine size rectangular windows, applies a Bayesian decision rule to detect minelike pixels and groups these pixels into objects from which features are extracted for binary classification. More recent papers have investigated the use of, e.g., Convolutional Neural Network (CNN)s and novel algorithms such as You Only Look Once (YOLO) (Redmon et al., 2016) to detect mines floating on the surface or in the water column (Munteanu et al., 2022). Chandran et al. (2002) have advanced the idea of Hyland and Dobeck (1995) by using features from higherorder spectra, reaching 90% detection accuracy by fusing bispectral and trispectral features from acoustic images. Another approach to detection and classification is done utilizing, e.g., SIFT, Harris, Shi-Tomasi and SURF feature detectors (Tueller et al., 2018), comparing the performance of Support Vector Machine classification using extracted features from a dataset of synthetic sonar images on different seabeds. Sun et al. (2022) have utilized Dual Path Vision Transformer Network (DP-ViT) to detect targets in forward-look sonar and SSS, but not for mine detection or classification.

Hożyń (2021) conducted a review of this research topic providing a comprehensive cross-section of the research situation from 1997 to 2020. The paper includes a background of mine warfare, MCM operations and sonar equipment but focuses on Computer Aided Detection (CAD), Computer Aided Classification (CAC) and Automatic Target Recognition (ATR) algorithms. Of the listed 22 research papers, 8 utilize classical image processing techniques, 5 use conventional ML and 9 use deep learning techniques. The identified tasks or applications are classification and detection, or both. Object detection aims to detect all instances of objects from one or several known classes (Zhao et al., 2019; Amit et al., 2021) while classification refers to correctly classifying images or image objects into known classes.

Seven of the Deep Learning (DL) papers reviewed by Hożyń (2021) from year 2000 onward are displayed in Table 1, where D denotes detection and C denotes classification. As the table shows, Gebhardt et al. (2017) have studied the use of deep Neural Network (NN) in detection of minelike objects with side-scan sonar imagery, reaching an accuracy of 98% with a modest-size network and their respective dataset. The dataset used was mostly synthetic, as described in previous work done by Barngrover et al. (2015). The use of CNN and transfer learning has been researched by McKay et al. (2017) with results indicating that impressive accuracy can be reached with transfer learning, with deeper models performing better than smaller

Table 1 Earlier research utilizing DL techniques for classification and object detection

Research	Task	Data	Methods	Performance
Yao et al. (2002)	C	40 kHz and 80 kHz CSS datasets	Small NN	91.94%–95.83% accuracy
Fei et al. (2015)	C	Atlas Elektroniks dataset, $n = 210$	KNN, SVMG, PNN	87.62% for PNN
Williams (2016)	C+D	Custom CMRE SAS datasets	CNN	0.887–0.979 AUC
McKay et al. (2017)	C	CSS datasets	VGG16, VGG19, Alexnet (Simonyan and Zisserman, 2015; Krizhevsky et al., 2017)	90.3%, 98.8%, 98.8%, respectively
Gebhardt et al. (2017)	D	Semi-synthetic (Barngrover et al., 2015)	Deep NN	98% accuracy
Wu et al. (2019)	C	Custom SSS data from Fujian, China	Encoder-Decoder	Mean accuracy 77.05%
Thanh Le et al. (2020)	D	Custom DSTG SSS dataset (Chapple, 2009)	Gabor NN	79.93% accuracy

equivalents. Wu et al. (2019) have used an encoder-decoder structure to capture the context of an SSS image, which is then processed with CNN, producing advantages and the ability for real-time processing.

Table 1 shows that the research and results share similarities, but are not directly comparable due to the lack of a common dataset and baseline models. For example, Fei et al. (2015) have used K-Nearest Neighbor (KNN), Support Vector Machine with Gaussian kernel (SVMG) and Probabilistic Neural Network (PNN), where classical methods function as a baseline. The research closest to the methodology of this paper is the research by McKay et al. (2017), using VGG and AlexNet (Simonyan and Zisserman, 2015; Krizhevsky et al., 2017), but data-wise Thanh Le et al. (2020) is the only one with operational Defence Science and Technology Group (DSTG) data from Australian armed forces. Two studies utilize Coastal Systems Station (CSS) data, while some custom datasets such as Centre for Maritime Research and Experimentation (CMRE) SAS datasets are also used. The resulting accuracies in detection and classification range from 77% to 98%, depending on the data and methods selected. The reported metrics also differ, e.g. Williams (2016) report Area Under the Curve (AUC) while others report accuracy percentages. It is noteworthy that lower accuracies are systematically related to more operational SSS data acquired in seabed survey missions.

None of the studies represented used data from the Baltic Sea area. Instead, a case study by Kunde and Schneider von Deimling (2016) suggests the use of a ship-born high-frequency multibeam sonar in shallow waters for ammunition and mine detection, according to results obtained in the southern Baltic Sea. Petillot et al. (2006) have proposed a framework for evaluating underwater mine detection with augmented reality, using also data from the Baltic Sea. The data has been collected with similar Klein 5 000 sonar, but the focus of the study is to introduce a framework to evaluate either a human operator or an ATR algo-

rithm. Petillot et al. (2006) also highlight that the Baltic Sea area can be tedious for mine hunting operations as some areas are cluttered with mine-like objects. Peine and Brecht (2005) have studied the detection of buried objects in the southern Baltic Sea with penetrative sonar echoes.

To the author's best knowledge, no public research exists on applying ML in mine detection in the littoral of the Baltic Sea using real-world data from mine hunting operations and exercises. Additionally, the use of a transformer architecture and MoE for mine detection and ATR algorithms remains unexplored in public research. The challenging conditions prevalent in the Finnish Archipelago and its surrounding waters have hindered the effective application of previous ATR algorithms, which has been driving the initiative and motivation for this research in the first place.

2 Experimental methodology and design

The benefit of automated mine detection and classification has been highlighted by Schwarz (2014) as enabling operators to focus on the most prominent targets and other tasks besides detection and classification. A reliable classification and detection system can reduce the time spent on MCM operations to a significant degree. It can also be argued that a detection system functions independently from human factors such as fatigue. Prolonged MCM operations would require operators to execute their tasks for days at a time without having time for proper recuperation. Such a work tempo is prone to induce errors in the workflow.

This work focuses on classifying the imagery produced by the Klein 5 500 SSS (Armstrong et al., 1998) in the exercises and operations of the Finnish Navy MCM units. A working classification model can boost the operator's workflow by highlighting objects of interest for the operator in object detection manner for easy distinction between mines and other objects on the seabed.

The methodology consists of transfer learning with a very deep CNN architecture VGG16 as well as VGG19 (Simonyan and Zisserman, 2015) and a vision transformer with an attention mechanism (Vaswani et al., 2017; Dosovitskiy et al., 2021). By leveraging diverse models, any patterns or features identified as areas of interest can be validated across different architectures, reducing the risk of model-specific biases or errors. Additionally, the approach enables the exploration of complementary strengths. Leveraging diverse models helps validate patterns across architectures, reducing model-specific biases and errors. Employing varied models facilitates flexibility in addressing specific challenges or constraints. For instance, one model may perform better in scenarios with limited data or noisy inputs, while another may offer superior performance in capturing fine-grained details. The separate model performances can be leveraged together to amplify the predictive performance with a MoE approach (Masoudnia and Ebrahimpour, 2014) if the models display strengths in different areas of the problem.

CNN contributes to a rising level of classification accuracy in the maritime environment (Baird et al., 2020). However, CNN needs plenty of data in order to achieve an acceptable level of accuracy. One solution to tackle the problem of data scarcity is transfer learning (Hutchinson et al., 2017). Transfer learning can transfer knowledge between domains due to the ability generalizable feature extraction. In this case, all deep learning models were pre-trained with ImageNet (Krizhevsky et al., 2012) data, enabling transfer learning to new classifier models.

Finally, as three different deep learning methods were utilized for the classification task, the resulting models are also tested as a MoE (Yuksel et al., 2012; Masoudnia and Ebrahimpour, 2014) to enhance the results by leveraging the strengths of all models in cohesion. According to author's best knowledge, MoE has not been used in this manner in previous research.

There are three key components: establishing a baseline with classical methods, comparing novel deep learning methods to classical performance, and testing the combination of these models in MoE framework. The data is pre-augmented before being fed to any of the architectures, producing classifier models that separate the data samples to the original four classes as well as a binary classifier that separates mines from other instances.

2.1 Sonar image dataset characteristics

The data provided by the Finnish Navy consist of a collection of 1 406 images. These images are not the raw data from the SSS system, but from a database that gathers the seabed mapping and geospatial data of objects. The image dataset was produced by exporting PDF reports with an image of the area in question: the raw data is not available as such.

The dataset contains images from four different classes as summarized in Table 2. The dataset is imbalanced, i.e. is unevenly distributed between different classes, which makes the ML solutions prone to minimize the error related to the dominant classes. 107 of the samples were disregarded after manual inspection, as the original set included duplicated instances and images so blurred or dim that recognizing any perceivable object or pattern was not possible by a human operator. Hence, the dataset used to training consists of 1 299 images.

Table 2 Number of samples in original and cleansed data sets

Class	Number of images	Training set
Mines	205	178
MILCO	406	383
Rocks	711	663
Wrecks	84	75
Total samples	1 406	1 299

The images were extracted from PDF images, with the size measuring 193 mm × 179 mm (182 mm × 163 mm for MILCO class images), and a resolution of 300 DPI. The scale varies for different classes: for the mine, rock, and MILCO classes, the scale is 1:150, while for the wreck class, it is 1:200. Consequently, the area covered by the images of wrecks is approximately 38 m × 35 m, whereas for other classes, it is 27 m × 24 m. The images were converted to JPG format and scaled to a smaller 224 × 224 size to enable quick processing with ML algorithms.

Figure 2 displays good-quality samples from the data that are easier to interpret. However, the distinction between a rock, a mine-like object and an actual mine is not intuitively clear. Figure 3 displays lower-quality samples that are more difficult to interpret. The images were extracted from an ArcGIS-based system that displays layered spatial information such as map layers; the background colour is from a certain map layer that has been set during the extraction of the image. The color composition of the images alternates between grayscale and sepia colourmaps, depending on operator preference during the initial analysis and extraction of the data. Additionally, sonar data was available only from surveyed areas and may contain images combined from several survey missions. This lead to discrepancies in the color scheme with sharp edges and different colors such as the blue base map shown in Figure 3. Furthermore, many images feature traces of transformation lines, attributable to the amalgamation of scanned measurements.

The distinction between classes is considerably more blurred with low quality samples in comparison with better-quality samples. For example, the left-most "Mine" class image in Figure 3 does not include a detectable mine from the image angle, even though there are known mines

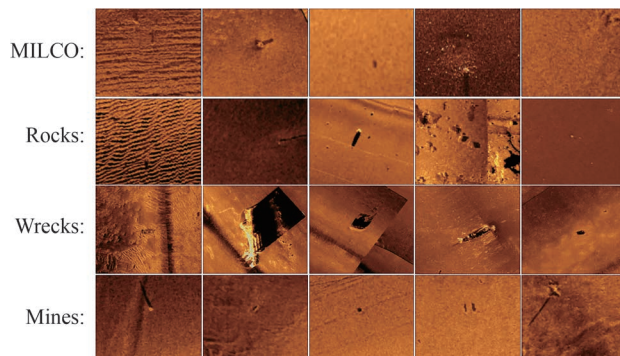


Figure 2 A sample of good quality examples of different classes in the collected dataset. The rows are in order: mine-like objects, rocks, wrecks, mines

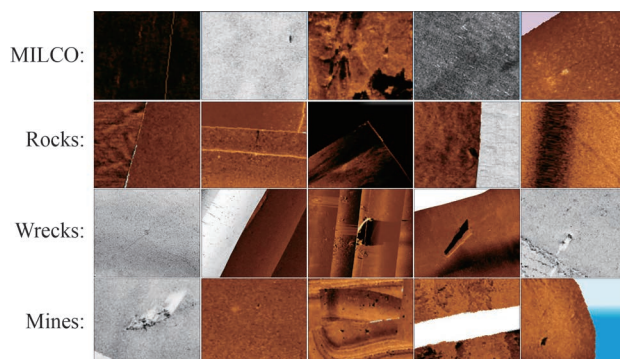


Figure 3 A sample of poor quality examples of different classes in the collected dataset

located in the area of the image hidden in the debris.

For training, validation and testing the data was evenly split into ten folds over all classes for all models. For each fold, approximately 75% of the data was allocated for training purposes, about 15% for validation, and the remaining 10% for testing.

2.2 Data augmentation

As the original dataset consists of a very small number of samples, the amount of data was increased via data augmentation (Perez and Wang, 2017) which is typical for pattern recognition tasks. The chosen data augmentation technique expands the training dataset to approximately eight times the size. Initially, horizontal and vertical flips were applied. Subsequently, the images were rotated by 90°, followed by another round of horizontal and vertical flips. To introduce variability, random translations of up to 20% of the image dimensions are applied, with color interpolation employed to fill in the additional areas created by the translations.

2.3 Machine learning framework

2.3.1 Training

The training phase was divided into three scenarios:

Case 1) classification with all four classes, Case 2) combining MILCO and mine classes to create a three-class classification task and Case 3) leveraging MoE and the best expert models from two previous approaches. Combining MILCO and mine classes serves the operational objective as MILCO targets are unconfirmed and treated as mine detections until further investigative actions have been carried out. To obtain an unbiased estimate of the model performance given the limited dataset, training was performed using 10-fold cross-validation with Stratified K-Fold (SCV) (Prusty et al., 2022). This methodology creates a unique training and validation set for each fold. SCV ensures the generalization of the model by maintaining the correct class distribution in each fold, addressing the imbalanced dataset problem.

The number of epochs was tested with an early stopping criteria approach and a fixed number approach, where the fixed number produced better results. The convergence is analyzed from the training and validation error, which showed that both models reached a plateau at respective training time; this was confirmed by comparing with the results from early stopping criteria.

2.3.2 Baselines: Support vector machines and random forest

As displayed by Hożyń (2021), many research papers use SVM as a baseline for performance comparison. To induce compatibility with previous research, an SVM (Boser et al., 1992) solution is implemented as a baseline, along with a Random Forest (RF) (Breiman, 2001). VGG16 architecture is used as feature extractor for both of the baseline models to enable the use of the classical ML frameworks.

2.3.3 Transfer learning with VGG16 and VGG19

The CNN models are pre-trained VGG16 and VGG19 (Simonyan and Zisserman, 2015) models that were fine-tuned with the SSS dataset, as done in McKay et al. (2017). The classifier for both architectures comprises of two dense layers with 256 and 128 nodes respectively. A dropout layer with a 10% probability of exclusion is placed between the two dense layers. Dropout layers help to reduce overfitting. The output is normalized into a categorical probability distribution with a Softmax activation function. The optimizer used to update the weights was Adam (Gupta et al., 2021), and the error was calculated with categorical cross entropy (Zhang and Sabuncu, 2018).

2.3.4 Transfer learning with vision transformer

The ViT model is a pre-trained ViT-B/16 (Dosovitskiy et al., 2021) model that is fine-tuned with SSS dataset. Unlike traditional CNNs, ViT employs a transformer-based architecture, originally designed for sequential data such as natural language processing, to process image data.

The ViT-B/16 model, characterized by its B variant with a patch size of 16×16 , is notable for its ability to capture long-range dependencies within images through self-atten-

tion mechanisms. Model divides the input image into fixed-size patches, which are then linearly embedded into lower-dimensional representations suitable for processing by transformer layers. The subsequent transformer encoder processes these embeddings, allowing for interactions between patches and making it easier to understand an image as a whole.

One of the strengths of the ViT models lies in the ability to effectively capture global contextual information within images, enabling them to discern intricate patterns and relationships across different regions (Maurício et al., 2023). This capability makes it particularly suitable for applications such as the SSS dataset, where understanding the broader context of an image can provide valuable insights into regions of interest, facilitating faster and more efficient localization of relevant features.

However, it is important to acknowledge certain limitations inherent in the ViT models. Vision transformers, including ViT-B/16, are notably data-hungry models, often requiring large-scale datasets for robust results. In the literature (Maurício et al., 2023), ViT models have demonstrated good performance, especially in complex dependency context learning, but are computationally expensive for resource-constricted environments. Conversely, traditional CNNs are good due to their translation invariance, local feature extraction, and easy pruning and distillation for performance optimization (Maurício et al., 2023).

To mitigate the risk of overfitting and enhance model generalization, a dropout layer with a 10% probability of exclusion is incorporated into the ViT-B/16 architecture, analogous to the approach adopted for the VGG models.

2.3.5 Mixture of experts

In a simple MoE framework, the output of several ML models or NNs is combined to enhance the predictive capability of the application (Masoudnia and Ebrahimpour, 2014; Yuksel et al., 2012). In MoE classification model, the expert softmax outputs are weighted with a gate that creates a weighted sum of the expert outputs with regard to the original input, defining regions where the individual experts are trustworthy (Yuksel et al., 2012). In this case, the gate is trained with either three or four experts, ViT, VGGs and SVM, according to Algorithm 1. The $\hat{Y}$ denotes the set of softmax output vectors $\hat{y}$ predicted by the trained models $m \in \mathcal{M}$, $y \in Y$ are actual one-hot encoded labels, g_θ is the gate network, $\hat{g}$ denotes weighted predictions and X is the dataset. The output weights for each model $m_n \in \mathcal{M}$ denoted as $[g_0, \dots, g_n \leftarrow g_\theta(x)]$ are optimized in a training loop with Adam optimizer and cross-entropy loss between the prediction and true label.

The MoE gate uses VGG16 architecture as feature extractor that was combined with three dense, linear heads that produce the model specific outputs for each category, i.e., class specific output values from each head. In essence,

Algorithm 1 Mixture of Experts Training

```

1: Initialize parameters  $\theta$  for gate network  $g_\theta$ 
2: Load dataset  $X$ , true labels  $Y$ , and expert models
    $\mathcal{M} = \{m_0, m_1, \dots, m_n\}$ 
3: Set the number of epochs  $N$  and learning rate  $\eta$ 
4: for epoch = 1, 2, ...,  $N$  do
5:   for all  $(x, y) \in (X, Y)$  do
6:      $\hat{Y} \leftarrow \emptyset$ 
7:     for all  $m_j \in \mathcal{M}$  do
8:        $\hat{y}_j \leftarrow m_j(x)$ 
9:        $\hat{Y} \leftarrow \hat{Y} \cup \{\hat{y}_j\}$ 
10:    end for
11:     $[g_0, g_1, \dots, g_n] \leftarrow g_\theta(x)$ 
12:     $\hat{g} \leftarrow \sum_{j=0}^n g_j \odot \hat{y}_j$ 
13:     $\mathcal{L} \leftarrow - \sum_{i=1}^C y_i \log(\hat{g}_i)$ 
14:     $\theta \leftarrow \theta - \eta \nabla \theta \mathcal{L}$ 
15:  end for
16: end for

```

the MoE gate learns to provide the mixture with input specific weights to enhance the classification by leveraging strengths of different models.

For clarity, the different MoEs are denoted as MoEv3 for the mixture of VGG16, VGG19 and ViT, whereas the four expert model with additional SVM is denoted MoEv4.

2.4 Classification and detection in practice

The trained models are used to classify, i.e., predict the label of a sample, after which Grad-CAM (Selvaraju et al., 2017) is employed to highlight the activation zones in the predicted images. Additionally, bounding boxes are drawn around the regions of interest (ROIs) by identifying the areas in the predicted images where VGGs and ViT detect potential details. A bounding box, in this context, is a rectangular border that encloses the region of an image where an object of interest is detected, thereby allowing for localization.

In the case of VGG16 and VGG19, Grad-CAM is used to determine the last convolutional layer activation map, highlighting the areas with the highest activations. For ViT, the average attention matrix across all heads is utilized to identify the regions with the most significant attention. The key area in both the activation map and attention matrix are selected based on the most activated regions. The final bounding box with the localized object of interest is then determined by identifying a region that encompasses both

of these key points, forming a rectangle that includes the areas identified by both models. This method prioritizes regions of high attention or activation, assuming they correspond to object locations.

While the current method of selecting bounding boxes offers a systematic approach to object localization, it's crucial to acknowledge its limitations. A primary concern arises when attention mechanisms produce noisy outputs, failing to pinpoint definitive spots of interest. In such cases, bounding boxes may inadvertently cover extensive areas, even spanning the entirety of the image. This issue highlights the necessity of enhancing model robustness to handle more challenging scenarios where clear solutions are elusive.

3 Experimental results

3.1 Classification performance

The performance of trained models is evaluated across two classification scenarios and two approaches as explained in subsubsection 2.3.1:

- Four class classification with individual models
- Three class classification with individual models
- MoE for both previous cases

The performance is evaluated by averaging accuracy, precision, recall, and F1-score over class-specific values, respectively. Accuracy is calculated as true positives (TP) out of all samples (n) in the particular class (i). Precision for each class is calculated as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

where FP denotes false positives, while recall for each class is calculated as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

where FN denotes false negatives. F1 score for each class is calculated as the product of precision and recall divided by their sum. The averaged metrics, i.e., the sums of metrics divided by number of classes, are displayed in Table 3, where the 4-class and 3-class classifier results are displayed consequently.

There are two MoE versions for both tasks, with three experts, i.e., VGG16, VGG19 and ViT, as well as a four expert version with additional SVM. These are denoted according to the number of experts as MoEv3 and MoEv4, respectively. As highlighted in Table 3, the MoE approach reaches highest performance metrics in all cases. As such, the MoE approach succeeds in leveraging the model-wise differences in the ability to correctly classify dataset samples.

Table 4 shows the overall accuracy for all models in

Table 3 Class-wise mean accuracy, precision, recall and F1 results for all classifier models

Case	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
4-class	RF	26.73	54.86	38.03	37.57
	SVM	34.12	52.77	55.73	43.90
	VGG16	34.31	52.22	57.56	49.97
	VGG19	31.00	49.53	52.24	45.92
	ViT	41.78	56.32	60.20	57.59
3-class	RF	33.34	48.50	48.55	48.52
	SVM	46.07	64.47	67.84	62.84
	VGG16	46.69	60.44	66.90	63.20
	VGG19	44.24	61.38	63.55	61.27
	ViT	51.74	66.16	70.83	68.04
MoE	4-class MoEv3	45.20	63.29	58.75	60.70
	4-class MoEv4	47.50	65.84	60.64	62.61
	3-class MoEv3	56.21	75.86	69.16	71.91
	3-class MoEv4	54.78	75.05	67.82	70.75

Table 4 Overall accuracy for all models

Overall accuracy	RF (%)	SVM (%)	VGG19 (%)	VGG16 (%)	ViT (%)	MoE (%)
4-class MoEv4	60.12	51.42	49.73	43.57	59.74	67.74
3-class MoEv3	57.66	63.28	64.90	61.20	70.59	73.29

both classification tasks, where the overall accuracy is calculated by dividing all true positives by the total number of samples as $\frac{\sum_i \text{TP}_i}{\sum_i n_i}$.

In total, the results reflect the difficulty of the classification task with this particular dataset. In previous research, as displayed in Table 1, accuracies in vicinity of 80% were expected for such a real-world operational data set. The difficulty is further amplified by the quality of data, as it was not received in a raw image format, and the quantity that totaled 1299 samples shown in Table 2.

However, the overall accuracy of 73.29%, with 75.86% precision and 69.16% recall of the 3-class MoEv3 shows that even with a small dataset, leveraging a number of deep learning methodologies promises adequate results.

3.2 Explainability outcomes

Figure 4 illustrates examples from each image class. Image (a) shows the original image, (b) visualizes the ViT attention map overlaid on the original, (c) displays the VGG Grad-CAM activation heatmap with the original image on top, and (d) highlights the area of interest with a rectangle, marking the highest common activation between (b) and (c).

The analysis focuses on identifying areas of interest rather than classifying the images. across the four examples, one per class. For the Mine and MILCO image examples, the common area of interest is smaller when both methods highlight region in the same area. Conversely, in the rock and wreck examples, when the methods activate on different regions, the common area of interest becomes larger.

When using MoE in the way it was utilized in this paper, as weighing the experts outputs in a multi-class setting with regard to the input, the explainability can be achieved in two manners: the first option is to first classify the object and select the model that had the highest softmax output for that particular class, and then use that particular model to create the GRAD-CAM activation heat map or attention heat map to highlight the area of interest. The alternative way is to combine the GRAD-CAM activation heat maps from VGG16, VGG19 and the attention heat map from ViT and visualize these as a compos-

ite image. With the latter method, there exists the possibility of destructive inference, if the models would have near opposite heat maps, which would have to be taken into consideration.

4 Discussion

The numerical results of this research were modest in comparison with earlier research. This is attributed to the quality and quantity of data, as similar methods have shown more promise with other data sets. However, the used data equals the data currently used for post-operation analysis for the same, manual task. Hence, the result highlights the compatibility of current systems and procedures to adapt AI into mine hunting practices in the Baltic Sea region.

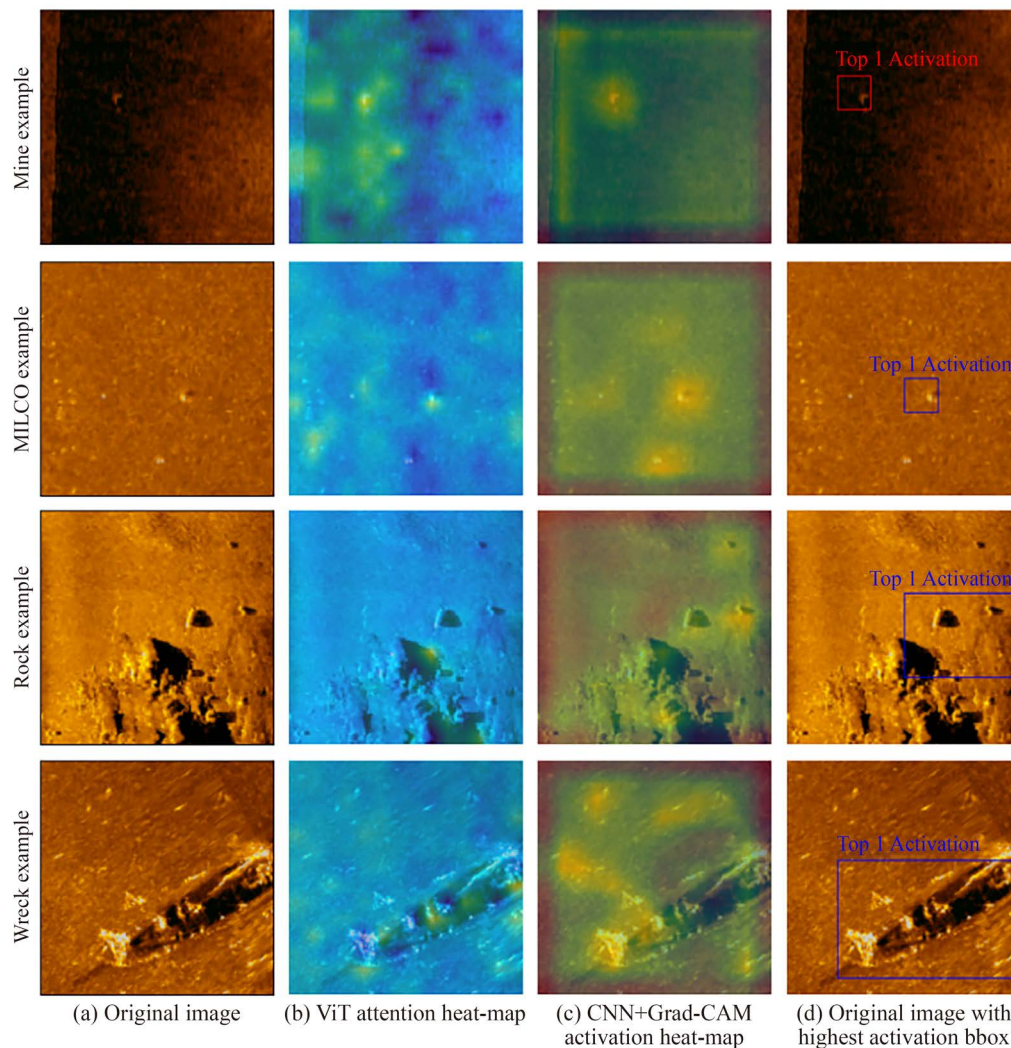


Figure 4 Illustrated example of explainable object detection from an original sample to highest activation mapping: (a) The original image. (b) ViT model mean attention map across all heads. (c) CNN GradCAM attention heat-map with the located object. (d) Pinpointed area of interest within the image

The most interesting result is the application of MoE methodology in improving the initial results. The applied method manages to leverage the strengths of multiple trained expert models to improve the results by a considerable margin.

It can be argued that the difficulty of the data set inhibited further improvement or would require a more suitable feature extractor. In the current setting, the VGG16 feature extractor functions well, but can also bottleneck the margin of improvement to its shortcomings that are visible in VGG16 results.

Likewise, the current MoE structure is computationally intensive, as it uses three deep inference models and the gate for weighting the expert outputs. This can be mitigated by training the MoE to function as a model selector instead of weighting the computed outputs. In this proposed setting, the MoE gate would propose the expert based on the input, after which the chosen expert would classify the particular sample. This approach is left for future research.

In future work the training dataset is intended to be expanded with synthetic data created with Stable Diffusion (Li et al., 2024). This is expected to increment mine detection accuracy and coverage in different maritime environments. Moreover, the research could focus on fortifying models to navigate such complexities and mitigate the risk of faulty detection. By addressing these challenges, the efficiency and accuracy of object localization can be significantly improved. The cost function used in training is to be weighted to reflect the anticipated cost or risk of particular classifications, in this case, false negatives.

Despite these limitations, the current methodology presents viable and novel means of automating and accelerating the processing of SSS images, paving the way for enhanced analysis and interpretation. For operational deployability, the resulting MoE classifier could be used in operational systems by slicing the SSS data into suitable, slightly overlapping frames for classification and region of interest detection, aiding in real-time or post-operation data analysis.

Further improvement of model performance can enhance the area of interest localization by enabling more accurate and precise feature extraction, better handling of noise and variability in the data, and increased robustness to complex patterns or occlusions.

5 Conclusion

In this article transfer learning with VGG16, VGG19 as well as state-of-the-art Vision Transformer were compared in a classification task in identifying seabed objectives from SSS images. The study highlights the difficulty of such data, where even sophisticated state-of-the-art deep learning methodologies do not reach remarkable results by

themselves. The intricate differences in pattern recognition between the models were successfully exploited with a novel MoE framework, which managed to improve the results by a significant margin: the MoE improvement compared to the next best single model was approximately 3 percentage points. With MoE and three deep learning expert models, the highest overall classification accuracy for a 3-class classification task between mines and mine-like objects, rocks and wrecks was 73.29%.

Additionally, the explainability and operator support were researched through Grad-CAM activation heat map and attention heat map utilization using the trained deep learning models, which can be combined with the MoE framework to reach both classification accuracy as well as visual operator support in detecting regions of interest when processing SSS imagery.

Funding This research was funded by MATINE (Finnish National Defence Scientific Advisory Board), Grant Number VN/15392/2022-SAAP/5 in collaboration with the University of Turku. Authors from the University of Turku have also received a research grant from the Department of Computing.

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

- Amit Y, Felzenszwalb P, Girshick R (2021) Object detection. Cham: Springer International Publishing, 875-883
- Armstrong A, Huff L, Glang G, Guenther G, Eisler T, Riley J, Perez S (1998) New technology for shallow water hydrographic surveys. *International Hydrographic Review* LXXV(2): 27-41
- Baird Z, McDonald MK, Rajan S, Lee SJ (2020) A CNN-ISTM network for augmenting target detection in real maritime wide area surveillance radar data. *IEEE Access* 8: 179281-179294
- Barngrover C, Kastner R, Belongie S (2015) Semisynthetic versus real-world sonar training data for the classification of mine-like objects. *IEEE Journal of Oceanic Engineering* 40(1): 48-56. <https://doi.org/10.1109/JOE.2013.2291634>
- Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT'92*, New York, 144-152
- Breiman L (2001) Random forests. *Machine Learning* 45: 5-32. <https://doi.org/10.1023/A:1010933404324>
- Brocki L, Chung NC (2023) Class-discriminative attention maps for vision transformers. *ArXiv abs/2312.02364*
- Broughton B (1998) The (R)evolution of mine countermeasures. *U.S. Naval Institute Proceedings* 124(5): 1143
- Chandran V, Elgar S, Nguyen A (2002) Detection of mines in acoustic images using higher order spectral features. *IEEE Journal of Oceanic Engineering* 27(3): 610-618. <https://doi.org/10.1109/joe.2002.1040943>
- Chapple P (2008) Automated detection and classification in high-resolution sonar imagery for autonomous underwater vehicle operations. Technical Report DSTO-GD-0537, Maritime Operations Division, Defence Science and Technology Organisation

- Chapple PB (2009) Unsupervised detection of mine-like objects in seabed imagery from autonomous underwater vehicles. *OCEANS* 2009, 1-6
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929v2*. *arXiv:2010.11929v2* [cs.CV]
- Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, Ranjan R (2023) Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* 55(9): 1-33. <https://doi.org/10.1145/3561048>
- Fei T, Kraus D, Zoubir AM (2015) Contributions to automatic target recognition systems for underwater mine classification. *IEEE Transactions on Geoscience and Remote Sensing* 53(1): 505-518. <https://doi.org/10.1109/TGRS.2014.2324971>
- Flemming BW (1976) Side-scan sonar: A practical guide. *The International Hydrographic Review* 53(1): 65-92
- Gebhardt D, Parikh K, Dzieciuch I, Walton M, Vo Hoang NA (2017) Hunting for naval mines with deep neural networks. *OCEANS 2017-Anchorage*, 1-5
- Gorman R, Sejnowski TJ (1988) Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks* 1(1): 75-89. [https://doi.org/10.1016/0893-6080\(88\)90023-8](https://doi.org/10.1016/0893-6080(88)90023-8)
- Gupta A, Ramanath R, Shi J, Keerthi SS (2021) Adam vs. SGD: Closing the generalization gap on image classification. *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 1-7
- Hansen FS (2025) Russian military thinking about the Baltic Sea and the Arctic. *DIIS Policy Brief*, Danish Institute for International Studies (DIIS)
- Hayes MP, Gough PT (2009) Synthetic aperture sonar: A review of current status. *IEEE Journal of Oceanic Engineering* 34(3): 207-224. <https://doi.org/10.1109/JOE.2009.2020853>
- Hibbert KR (2006) A need for systems architecture approach for next generation mine warfare capability. Master thesis, Naval Postgraduate School, Monterey, USA
- Hozýn S (2021) A review of underwater mine detection and classification in sonar imagery. *Electronics* 10(23): 2943
- Hutchinson ML, Antono E, Gibbons BM, Paradiso S, Ling J, Meredig B (2017) Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099*
- Hyland JC, Dobeck GJ (1995) Sea mine detection and classification using sidelooking sonar. In A. C. Dubey, I. Cindrich, J. M. Ralston, and K. A. Rigano (Eds.), *Detection Technologies for Mines and Minelike Targets*, Volume 2496, 442-453
- Kaskela A (2017) Seabed landscapes of the Baltic Sea: Geological characterization of the seabed environment with spatial analysis techniques. PhD thesis, University of Helsinki, Helsinki, Finland
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 25, 1097-1105
- Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60(6): 84-90. <https://doi.org/10.1145/3065386>
- Kunde T, Schneider von Deimling J (2016) Hydroacoustic detection of dumped ammunition in the ocean with multibeam snippet backscatter analyses. A case study from the 'Kolberger Heide' ammunition dump site (Baltic Sea, Germany). In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, EPSC2016-9168
- LeCun Y, Bengio Y (1998) *Convolutional networks for images, speech, and time series*. MIT Press, Cambridge, USA, 255-258
- Li W, Li C, Jiang H, Wang Y, Wu S, Wu Z (2024) High-resolution aerial image restoration with latent diffusion models. *2024 IEEE International Conference on Unmanned Systems (ICUS)*, 1874-1878
- Mansell T, Skinner D (1995) Application of Artificial Intelligence to decision making in Mine Countermeasures. Technical Report DSTO-TR-0279, Maritime Operations Division, Aeronautical and Maritime Research Laboratory
- Masoudnia S, Ebrahimpour R (2014) Mixture of experts: a literature survey. *Artificial Intelligence Review* 42: 275-293. <https://doi.org/10.1007/s10462-012-9338-y>
- Maurício J, Domingues I, Bernardino J (2023) Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences* 13(9): 5521. <https://doi.org/10.3390/app13095521>
- McKay J, Gerg I, Monga V, Raj R (2017) What's mine is yours: Pretrained CNNs for limited training sonar ATR. *OCEANS 2017-Anchorage*, Anchorage, USA, 1-7
- Melia T (1991) *Damn the Torpedoes: A Short History of U.S. Naval Mine Countermeasures, 1777-1991*. Contributions to Naval History, Naval Historical Center
- Munteanu D, Moina D, Zamfir CG, Petrea M, Cristea DS, Munteanu N (2022) Sea mine detection framework using YOLO, SSD and EfficientDet deep learning models. *Sensors* 22(23): 9536. <https://doi.org/10.3390/s22239536>
- Peine H, Brecht D (2005) Detection of objects buried in the seafloor. *Europe Oceans 2005*, Volume 1, 177-181
- Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv: 1712.04621*
- Petillot Y, Reed S, Coiras E, Bell J (2006) A framework for evaluating underwater mine detection and classification algorithms using augmented reality. Technical report, Heriot-Watt University
- Prusty S, Patnaik S, Dash SK (2022) SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology* 4: 972421
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788
- Schwarz M (2014) Future mine countermeasures: No easy solutions. *Naval War College Review* 67(3): 123-141
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618-626
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations*, 1-14
- Sun Y, Zheng H, Zhang G, Ren J, Xu H, Xu C (2022) DP-ViT: A dual-path vision transformer for real-time sonar target detection. *Remote Sensing* 14(22): 5807. <https://doi.org/10.3390/rs14225807>
- Thanh Le H, Phung SL, Chapple PB, Bouzerdoum A, Ritz CH, Tran LC (2020) Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. *IEEE Access* 8: 94126-94139. <https://doi.org/10.1109/ACCESS.2020.2995390>
- Tueller P, Kastner R, Diamant R (2018) A comparison of feature detectors for underwater sonar imagery. *OCEANS 2018 MTS/IEEE Charleston*, 1-6
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In I.

- Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Volume 30
- Williams DP (2016) Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. 2016 23rd International Conference on Pattern Recognition (ICPR), 2497-2502
- Wu M, Wang Q, Rigall E, Li K, Zhu W, Bo H, Yan T (2019) ECNet: Efficient convolutional networks for side scan sonar image segmentation. *Sensors (Basel, Switzerland)* 19(9): 48-56. <https://doi.org/10.3390/s19092009>
- Yao D, Azimi-Sadjadi M, Jamshidi A, Dobeck G (2002) A study of effects of sonar bandwidth for underwater target classification. *IEEE Journal of Oceanic Engineering* 27(3): 619-627. <https://doi.org/10.1109/JOE.2002.1040944>
- Yuksel SE, Wilson JN, Gader PD (2012) Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems* 23(8): 1177-1193. <https://doi.org/10.1109/TNNLS.2012.2200299>
- Zhang Z, Sabuncu MR (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, New York, USA*, 8792-8802
- Zhao ZQ, Zheng P, Xu ST, Wu X (2019) Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems* 30(11): 3212-3232. <https://doi.org/10.1109/TNNLS.2018.2876865>