Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Explainable Deep Learning to Classify Royal Navy Ships

BART BAESENS[1,4], AMY ADAMS[3], RODRIGO PACHECO-RUIZ[3], ANN-SOPHIE BAESENS[1], and SEPPE VANDEN BROUCKE.[2]

[1]Research Centre for Information Systems Engineering (LIRIS), KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium (e-mail: Bart.Baesens@kuleuven.be; AnnSophie.Baesens@yahoo.com)
[2]Department of Business Informatics and Operations Management, UGent, Tweekerkenstraat 2, 9000 Gent, Belgium (e-mail: Seppe.vandenBroucke@UGent.be)
[3]National Museum of the Royal Navy, HM Naval Base (PP66), Portsmouth, PO1 3NH, UK (e-mail: Amy.Adams@NMRN.org.uk; Rodrigo.Pacheco-Ruiz@NMRN.org.uk)
[4]School of Management, University of Southampton, 2 University Road, Highfield, Southampton, SO17 1BJ, UK

Corresponding author: Bart Baesens(e-mail: Bart.Baesens@kuleuven.be).

ABSTRACT We research how deep learning convolutional neural networks can be used to to automatically classify the unique data set of black-and-white naval ships images from the Wright and Logan photographic collection held by the National Museum of the Royal Navy. We contrast various types of deep learning methods: pretrained models such as ConvNeXt, ResNet and EfficientNet, and ConvMixer. We also thoroughly investigate the impact of data preprocessing and externally obtained images on model performance. Finally, we research how the models estimated can be made transparent using visually appealing interpretability techniques such as Grad-CAM. We find that ConvNeXt has the best performance for our data set achieving an accuracy of 79.62% for 0-notch classification and an impressive 94.86% for 1-notch classification. The results indicate the importance of appropriate image preprocessing. Image segmentation combined with soft augmentation significantly contributes to model performance. We consider this research to be original in several aspects. Notably, it distinguishes itself through the uniqueness of the acquired dataset. Additionally, its distinctiveness extends to the analytical modeling pipeline, which encompasses a comprehensive range of modeling steps, including data preprocessing (incorporating external data, image segmentation, and image augmentation) and the use of deep learning techniques such as ConvNeXt, ResNet, EfficientNet, and ConvMixer. Furthermore, the research employs explanatory tools like Grad-CAM to enhance model interpretability and usability. We believe the proposed methodology offers lots of potential for documenting historic image collections.

INDEX TERMS Convolutional Neural Networks; Deep Learning; Explainability; Digitised Archives; Image Classification; Royal Navy

I. INTRODUCTION

THE documentation and management of extensive digital or digitised archives pose a significant challenge for museums, archives, and historical collections worldwide. The field of Artificial Intelligence (AI) presents numerous novel applications that have the potential to revolutionise practices in digital archiving and collection documentation [1], [2]. In particular, as highlighted by [3], AI offers archival practitioners a framework to redefine, advocate, and delineate their expertise in digital archiving. The advent of deep learning (DL) techniques, coupled with advancements in graphical processing unit (GPU) acceleration has ushered in various innovative approaches for digital documentation.

DL techniques represent mathematical models inspired by the functioning of the human brain, designed to learn intricate, generalisable patterns from data. This understanding aims to unearth semantic meanings and reveal latent, intriguing connections among data elements [4]–[6]. The insights gained can be effectively leveraged for improved documentation, cataloging, clustering, classification, enrichment, labeling, or augmentation of structured and unstructured information, including digitised or digitally born texts, images, audio, and video materials.

This paper is based on the analysis of a unique dataset of digitised images of British Royal Naval ships obtained from the National Museum of the Royal Navy, and complemented with publicly available images. Each image was subjected to meticulous manual labeling into various predefined ship

types or categories, such as submarines, cruisers, destroyers, carriers, and more. Subsequently, deep learning convolutional neural networks (CNNs) were trained to autonomously recognise these classifications based on image attributes [7], [8].

The goal of the trained CNNs is to assist the National Museum in documenting, labeling, and organising more efficiently their extensive and ever-expanding archive of historic ships images. As such, and to enhance performance, various preprocessing techniques were explored, including image segmentation for automated removal of non-essential image components and image augmentation to generate new synthetic images from existing ones, thus enhancing generalisation performance. Recognising that CNNs are inherently complex, mathematical black-box models, we also shed light on their internal workings through explanatory tools. These visual explanations help clarify why a ship image was categorised into a particular class (e.g., cruiser or destroyer), aiding researchers on validation and labeling tasks, especially when the DL system exhibits uncertainty. This newfound clarity is also expected to contribute to more efficient image searches within the colossal image databases maintained by the National Museum, where many search queries are concise and often relate to a ship's categorisation (e.g., destroyer or cruiser).

From a scientific perspective, this paper makes several noteworthy contributions. First, utilising a unique set of labeled digital ship images, it introduces a robust and reproducible empirical methodology that combines cutting-edge image preprocessing with deep learning CNNs, enhanced by explanatory facilities for improved interpretation. Second, it empirically evaluates the interplay between image segmentation and augmentation as critical preprocessing activities to enhance the performance of the estimated deep learning image classification models. Lastly, the paper offers its Python and TensorFlow Keras-based code as open-source, with the potential for utilisation by collection, archive, or document managers in other digital collections settings. The insights gained are easily transferable to other image classification settings and digital collections.

The structure of this paper is as follows: Section II provides a comprehensive literature review of DL applications in documentation and ship classification. Section III details the origin of our ship image dataset, while Section IV covers data preprocessing activities. In Section V, we elaborate on the configuration and training of the deep learning models. Section VI presents the empirical findings, and in Section VII, we discuss methods to enhance the transparency and comprehensibility of the estimated deep learning models. Finally, Section VIII wraps up the paper.

II. LITERATURE REVIEW

Since deep learning is a relatively recent research discipline, with new techniques being continuously developed and/or perfected, some preliminary research has already been reported in the literature for documentation purposes. It's worth noting that deep learning techniques have demonstrated their prowess in handling substantial volumes of unstructured data. In what follows, we give some examples of previous research on using DL for classifying text, image, and/or audio data. Special emphasis is placed on convolutional neural networks (CNNs) and their application in ship image classification, as this forms the core focus of our study.

In the realm of text data, commonly applied deep learning (DL) techniques include transformers and bidirectional long short-term memory (Bi-LSTM) neural networks. [9] leveraged a Bi-LSTM deep learning neural network for annotating a dataset drawn from tourism and cultural heritage documents, including sources like Booking.com and TripAdvisor. In a comprehensive and unbiased evaluation of deep learning methods for text classification, [10] established that, on the whole, Bi-LSTMs were ranked as the top-performing approach, although their superiority over simpler methods like logistic regression was not statistically significant.

Convolutional Neural Networks (CNNs) represent the predominant deep learning (DL) technique for image analysis [8], [11]. A CNN operates as a feedforward neural network designed to extract image features by applying filters, also referred to as kernels or feature detectors, to the image. More specifically, each filter represents a particular image pattern, which is systematically traversed over the image's pixels, while convolution operations amalgamate the image input and the filter to generate a set of acquired features. As the network advances through its layers, the features evolve, becoming progressively more intricate and meaningful. Essentially, CNNs execute a hierarchical decomposition of the image, commencing with fundamental features such as lines, edges, contours, corners, and colors, before progressing to more complex elements like shapes (rectangles, circles, ellipses, etc.), and ultimately recognizing high-level concepts such as a submarine's fin, a destroyer's gun turret, or a carrier's flight deck in the deepest layers. Popular CNN implementations, listed in reverse chronological order, include ConvNeXt [12], ConvMixer [13], EfficientNet [14], DenseNet [15], ResNet [16], GoogLeNet [17], VGG [18], and AlexNet [19]. These CNN architectures typically vary in terms of their network architecture, such as the number of processing layers, the types of filters employed, the convolution operations, the training methods, and the estimated number of parameters, which can often extend into the millions.

All the previously mentioned CNN variants have undergone pretraining using publicly accessible datasets, such as the ImageNet database, which boasts a vast repository of 14,197,122 annotated images spanning 1000 distinct classes, each aligned with the WordNet hierarchy (e.g., goldfish, cowboy boot, broom, container ship, etc.) [20]. These learned representations or image features can subsequently be harnessed in a transfer learning framework for diverse image classification tasks, mirroring our approach in this paper. In other words, this entails the construction of a new CNN model by capitalizing on the pre-trained features from, for instance, a ConvNeXt or EfficientNet model, which was

originally trained on the ImageNet database. The model is then fine-tuned to adapt to the specific classification task at hand by incorporating additional network layers. With regards to our study, we opt for the most recent CNN variant, ConvNeXt, given its empirically demonstrated superior performance, as evidenced in [14]. In our subsequent analysis, we compare it against EfficientNet, ResNet and ConvMixer, all of which have also showcased commendable performance in prior research.

Ships can be categorised based on diverse sources of input data. One such example is acoustic signals captured through hydrophones, which record the radiated noise emitted by ships. These signals or audio streams can be readily transformed into spectrogram images, presenting a time-varying visual depiction of the frequency spectrum. Subsequently, CNNs can be employed for the analysis of these spectrogram images. [21] adopted this approach, utilizing CNNs including AlexNet, ResNet, and DenseNet, to classify ships as inbound or outbound based on audio signals collected from hydrophones. Another option involves the use of Synthetic Aperture Radar (SAR) data obtained from emitted radio waves. This data source was leveraged by [22], who employed CNNs to predict ship presence, position, length, and type. The utilization of optical remote sensing data for ship detection and classification has been explored and surveyed by [23]. In addition, [24] used satellite imagery data sourced from Kaggle to classify images as either containing ships or not, encompassing scenarios like open sea, clouds, or land. This task was tackled using both traditional AI methods and CNNs, including ResNet and DenseNet.

Images are a key source in the study of contemporary and historic naval ships, including in the study of shipwrecks as archaeological sites. In maritime archaeology the documentation of shipwreck sites currently relies heavily on the production and documentation of high definition (HD) and ultra high definition (UHD) imagery, including the use of 3D and 4D photogrammetry [25]. In some cases, the documentation of deep sea sites depends exclusively on the use of robotic generated imagery where conventional diving methods, such as diver based photography or acoustic 3D surveys from hull mounted survey vessels, cannot be utilised [26]. The use of automation and computational analysis has proven to be a significant driving force in the development of new research on such sites and opens another new door for the implementation of AI technologies in the research of these sites as described in [27] and [28].

With regards to contemporary ships and vessels [29] employed ResNet and AlexNet to analyse the publicly accessible Maritime Vessel (MARVEL) dataset. This extensive collection comprises 140,000 distinctly labeled maritime vessel images, spanning 26 diverse classes that encompass both civilian and military ships. In a comparative exploration, [30] introduced a ResNet extension and demonstrated its superior performance when contrasted with AlexNet, VGG, ResNet, and GoogLeNet. Their analysis encompassed a dataset comprising 8,932 images, categorizing ships into five classes, encompassing both civilian and non-civilian categories, alongside the MARVEL dataset. [31] also adopted CNNs for ship classification. Their approach commenced with the training of an AlexNet model, distinguishing between three classes: aircraft carriers, warships, and civilian ships. The dataset used for this phase comprised 250 images per class. Subsequently, a GoogLeNet model was trained to classify warships into subcategories, including coastal combat ships, shipyard transport ships, amphibious assault ships, submarines, and destroyers using a second data set containing 240 images for each class. The authors effectively showcased the efficacy of their method for ship image classification. [32] utilized a VGG model for a dataset of 2,400 images, classifying ships into four categories, encompassing both military and civilian vessels. They highlighted how data augmentation and fine-tuning of the VGG architecture contributed to improved model performance. [33] leveraged VGG, ResNet, DenseNet, AlexNet, and various other CNN variants for ship classification. Their self-collected dataset comprised 2,635 internet images, categorised into eight target categories, covering both civilian and non-civilian ships. Finally, [34] achieved success with AlexNet, VGG, and ResNet on a dataset consisting of 867 images, focusing on the classification of civilian ships into three categories.

Our study makes several contributions to the existing body of literature. Firstly, we adopt the recently introduced ConvNeXt method for the classification of military ships and provide a comparative analysis against a plain vanilla CNN, ResNet, EfficientNet and a ConvMixer model. Next, we conduct an in-depth exploration of the effects and interplay of data pre-processing techniques, including image segmentation and augmentation, on the performance of the CNN models under examination. Finally, our study goes beyond mere performance benchmarking by incorporating a visual explanation that elucidates the specific image elements upon which the estimated CNNs concentrate to make their classifications.

III. DATA COLLECTION

The bulk of the data set originated from the Wright and Logan collection held by the National Museum of the Royal Navy. Wright and Logan was a Portsmouth based photographer, who specialised in portraits of British Royal Naval or HMS (Her/His Majesty's ship) warships entering and leaving the key naval base in Portsmouth Harbour. All 3,533 images were black and white and taken between 1924 and 1998. Some example images are shown in Figure 1.

We enriched the data with 786 images scraped from https://www.naval-history.net/ and 1,324 images scraped from https://uboat.net/ using Python's Beautiful Soup package [35].

The unique image collection underwent a meticulous manual labeling process, facilitated by human input. Specifically, we engaged four labellers, who collectively dedicated approximately one man-month to the task by adding the ship name into the filename. During this process, images falling under certain criteria were excluded from the dataset. These

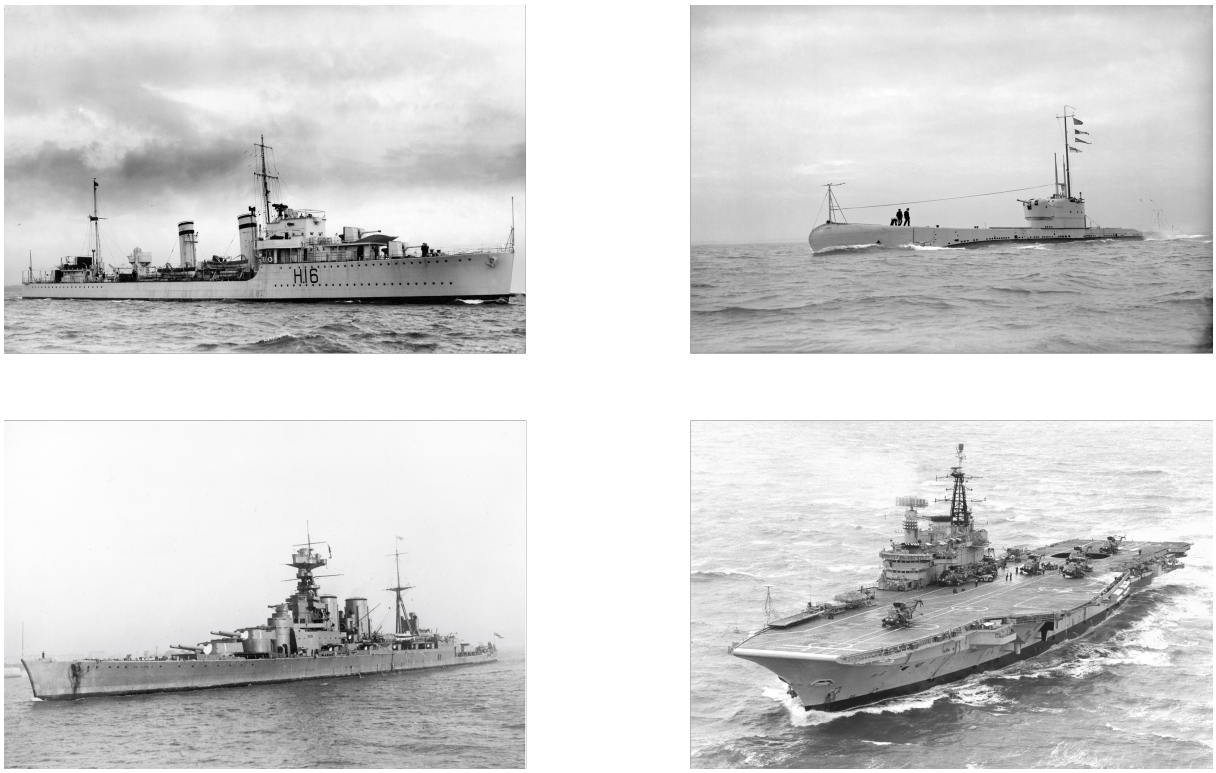


FIGURE 1: Example images of Wright and Logan collection (1924-1998). Top Left: HMS Daring destroyer (1932); Top Right: HMS Perseus submarine (1932); Bottom Left: HMS Hood battlecruiser (1918); Bottom Right: HMS Hermes carrier (1953).

included unclear images, those featuring multiple ships, images with intricate backgrounds (such as depictions of harbors, zeppelins, flying helicopters or jets, dense smoke clouds emanating from ship funnels, or images with extensive textual elements, like postcards), images portraying heavily damaged ships (such as HMS Vindictive following the Zeebrugge raid), and images offering glimpses of ship interiors. To facilitate the labeling procedure, we developed a Python web application, as illustrated in Figure 2. The labeling itself drew upon insights and guidance from curators at the National Museum, along with reference sources such as "Jane's Fighting Ships" [36], "Conway's All the Fighting Ships" books, and Wikipedia, as needed.

The data collection process presented a couple of noteworthy challenges. Firstly, ship names are reused in the Royal Navy. For instance, HMS Churchill served as a destroyer during the Second World War, but the same name was later also used for a nuclear submarine commissioned in 1970. Secondly, certain ship classifications, such as submarines and aircraft carriers, are relatively recognisable; however, early developments of these classes are more complicated. For example, the earliest carriers were typically just battleships with small runways constructed above the ship's forecastle as illustrated for HMS Barham in Figure 3. Furthermore, some ship types are subject to change, often being refitted for alternative purposes. An example of this transformation is the conversion of approximately 23 destroyers into type 15 frigates between 1949 and 1957 (e.g., HMS Relentless, HMS Ulster, and HMS Wakeful). The Wright and Logan collection includes pictures of the same ship both before and after refitting. Pennant numbers typically serve as a reliable indicator of a ship's type, with designations like "F" for frigates, "M" and "J" for minesweepers, "D," "H," and "R" for destroyers and "C" for cruisers. However, it's worth noting that their reliability is not perfect. For instance, HMS Ashanti was a destroyer with pennant number F51. Additionally, pennant numbers starting with "L" may indicate both a landing ship, like HMS Parapet L4039, and a destroyer, as exemplified by HMS Cossack L03, which also held the designations F03 and G03 at different points in time. Moreover, many ships lacked visible pennant numbers, making their identification a more intricate process.

In what follows, we elaborate on the data preprocessing, DL model configuration and training, and results. We note that the code and sample data for all experiments are publicly available on https://github.com/Macuyiko/royal-navy-ship-identification.

IV. DATA PREPROCESSING

To facilitate the training of deep learning models, we uniformly resized all images to a maximum width and height of 720 pixels, a resource-intensive task, using Python's PIL imaging library (https://pypi.org/project/Pillow/). This resizing was particularly demanding due to some of the

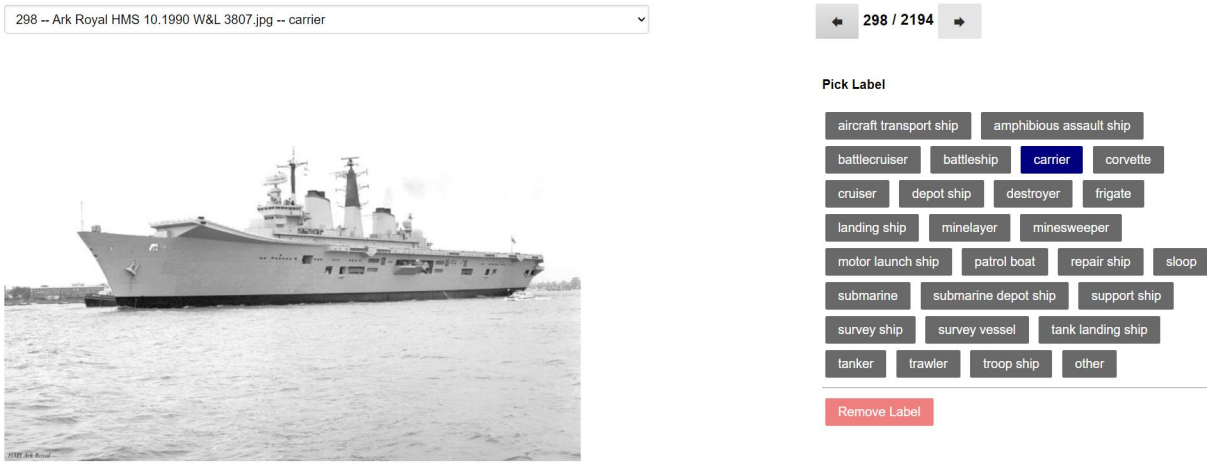


FIGURE 2: Illustration of Python labelling app picturing HMS Ark Royal (1990) being labelled as a carrier.

FIGURE 3: Plane flying off turret of HMS Barham ($\sim$ 1918) (Courtesy National Museum of the Royal Navy).

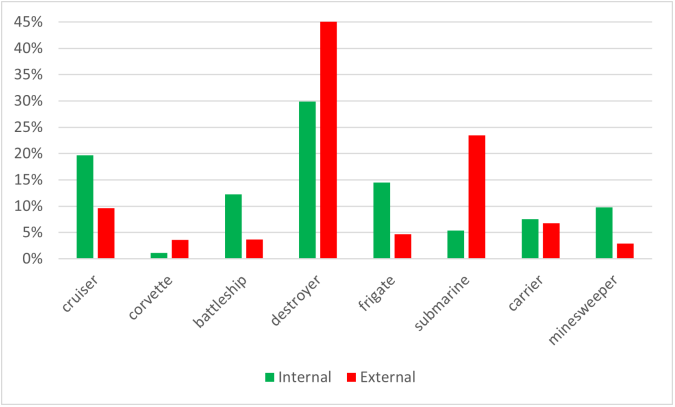


FIGURE 4: Histogram of target class distribution for internal and external data.

original TIFF files being quite large, with sizes reaching up to 35 megabytes, and occasionally, not well-formatted. Following a thorough examination of the class distribution and in consultation with the National Museum's curators, we devised deep learning models for classifying images into the following categories: battleship, carrier, corvette, cruiser, destroyer, frigate, minesweeper, submarine, and a residual category which was left out during training. The residual category included, among others, landing ships, repair ships, amphibious assault ships, sloops, survey vessels, trawlers, and depot ships. This category was omitted from the training due to the scarcity of observations, making any meaningful analytical discrimination unfeasible. The distribution of target classes for both the internal and external data is illustrated in the histogram presented in Figure 4, with destroyer being the most prevalent class and corvette the least represented.

In the process of training DL models, it is imperative to allocate a distinct, independent test set to facilitate model validation. This step is vital as it ensures that we obtain an impartial and fair assessment of model performance. In essence, all DL models undergo estimation on a training set, where various forms of data preprocessing can be explored. Subsequently, these models are evaluated using an entirely separate test set. In our case, we created our test set by randomly selecting 20% of the images obtained from the National Museum. As for the training set, we estimated models using the original remaining 80%. Additionally, we investigated the influence of augmenting the training set with externally obtained images on model performance.

To enhance the performance of the DL models, we implemented two image preprocessing techniques. The first involved the utilization of the Segment Anything (SAM) model, developed by Meta AI, accessible at https://segment-anything.com/, as introduced by [37]. SAM is an AI model renowned for its ability to efficiently remove irrelevant objects from images using a simple point-and-click mechanism. It stands as a fully pretrained and promptable AI model, offering zero-shot generalisation capabilities to

previously unseen images, making it a valuable asset for image classification. We harnessed SAM to automatically eliminate elements like the skyline, water, and harbor objects by applying click-based masks at the top, middle, and bottom portions of each image. This additional preprocessing step was designed to encourage the deep learning model to concentrate solely on ship characteristics during its classification process. The outcomes of applying SAM to images of HMS Agincourt, HMS Acasta, and HMS Repulse are portrayed in Figure 5. It's essential to note that while SAM filtering is beneficial, it may not always deliver perfect results, and traces of the background could persist in the processed images as can be seen in the figure.

Another preprocessing technique we explored was image augmentation. The fundamental concept behind this approach is to generate synthetic images derived from existing ones, thereby expanding the training dataset and affording convolutional neural networks more opportunities to discern and generalize meaningful patterns from a relatively limited set of labeled images. A variety of image augmentation operations can be considered, including but not limited to horizontal flipping, cropping, blurring, sharpening, and resizing. Other popular augmentation techniques involve introducing elements such as spatter, Gaussian noise, Gaussian blur, or even simulating fog within the image. It's worth noting that these augmentation operations are typically applied in combination to create a new, augmented image.

In our research, we used the well-established Albumentations Python library, accessible at https://albumentations.ai/, to facilitate fast and versatile image augmentation during the training process. Figure 6 presents an illustrative example using the destroyer HMS Capel. On the left, you can observe the original image, while on the right, there are four augmented images generated through these operations.

In our initial experiments, we observed that augmenting the segmented images often led to a degradation in performance. A closer examination unveiled that this decline was primarily attributable to several augmentations introducing additional non-white pixels, such as spatter, fog, blur, noise, within the previously masked or white areas resulting from the segmentation. To address this issue, two potential solutions were considered. The first approach involved applying augmentations to the images before performing the image segmentation during the model training process. Regrettably, this method proved to be excessively resource-intensive and non-scalable. Consequently, we adopted an alternative strategy, where we exclusively considered safe augmentations, such as horizontal flipping, sharpening, rotation, and image resizing. These augmentations were chosen for their non-interference with the white masks, ensuring that only the ship's structure underwent alterations. Figure 7 presents a clear example of these safe augmentations as applied to the battleship HMS Agamemnon. You can clearly see that the white masks remain and only the ship's corpus is undergoing changes.

V. DEEP LEARNING MODEL CONFIGURATION AND TRAINING

For our analysis, we employed a range of CNN variants, including a plain vanilla CNN model built from scratch, ConvNeXt, ResNet, and EfficientNet. The latter three models were employed within a transfer learning framework, wherein their filters and feature maps were pre-trained on the ImageNet dataset. Subsequently, additional layers were introduced, which were initially trained independently and fine-tuned on our ship image classification dataset. In the case of ResNet, we augmented the learned features with a fully connected output layer, often referred to as a dense head, which played a pivotal role in the final classification task. The classification was achieved through the utilisation of a softmax activation function [4]. For EfficientNet and ConvNeXt, we introduced a two-dimensional global average pooling layer, which calculated the average values of each feature map. These values were then channeled into a dense head, a structure illustrated in Figure 8.

Due to the highly parameterised nature of DL techniques, there's a susceptibility to fitting noise or idiosyncrasies within the data, a phenomenon often referred to as overfitting. To mitigate this issue, a commonly employed approach is the utilisation of dropout, which involves randomly deactivating specific network nodes during training, determined by a predefined dropout probability, as outlined in Table 1. Given the multiclass classification nature of our task, with eight distinct targets, all networks are inherently trained to maximize the log likelihood of the data, technically referred to as minimising a cross-entropy error objective [4]. However, it's worth noting that this objective function sometimes falls short in prioritising challenging-to-classify examples. This challenge becomes particularly evident in imbalanced class settings, as observed here with a notably dominant destroyer class and a rare corvette class. To address this concern, alternative objective functions like focal loss have been introduced in the literature [38]. In our preliminary experiments, we explored both the traditional cross-entropy loss and focal loss, employing default parameter settings for each of the DL techniques. The outcomes of these experiments are detailed in Table 1, highlighting the configurations that delivered the best performance.

All CNN models in our experiments were trained with the Adam (Adaptive Moment Estimation) optimiser responsible for determining the DL model parameters, such as filters and feature maps, by means of a gradient descent procedure. This process involves multiple sweeps, often referred to as epochs, through the training data. During each epoch, the optimiser takes downward steps on the error surface according to a predefined learning rate. Our chosen learning regime encompassed three distinct phases: three epochs with a learning rate of 0.01, followed by ten epochs with a learning rate of 0.0001, and concluding with 30 epochs using a learning rate of 0.000001. The rationale behind this schedule was to initiate the training with larger steps, promoting substantial error reduction in the early stages, and

FIGURE 5: Removing skyline, water and harbour objects from ship images using SAM. From left to right: HMS Agincourt (1913), HMS Acasta ($\sim$ 1912) and HMS Repulse ($\sim$ 1968).

FIGURE 6: Illustration of augmentation for HMS Capel (1942).

FIGURE 7: Illustration of safe augmentation for HMS Agamemnon (1906).

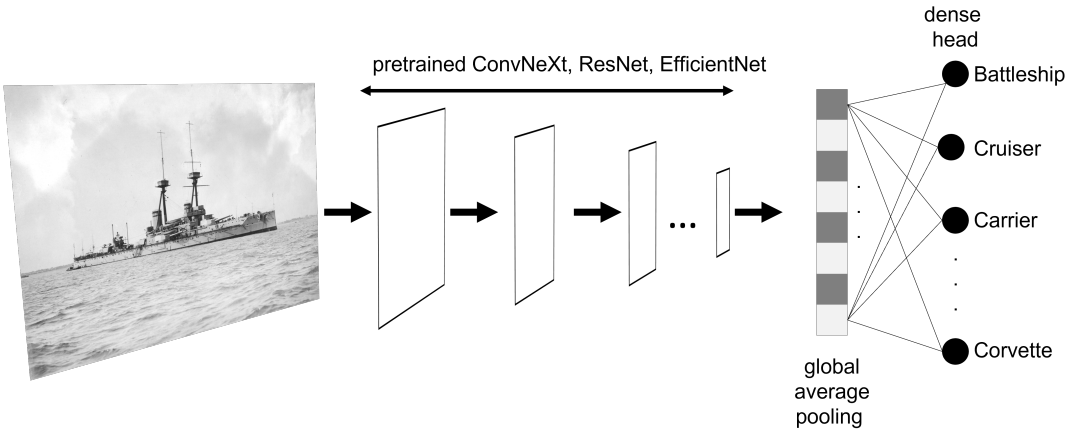


FIGURE 8: Illustration of pretraining (HMS Vanguard, 1913).

then gradually decrease the step size to avoid overshooting good (local) minima on the error surface. In scenarios where a pre-trained backbone model, such as ConvNeXt, ResNet, or EfficientNet, was incorporated, the initial three epochs involved freezing the backbone model. Only the dense head or fully connected output layer was actively trained during this period. Subsequently, the complete model was unfrozen, allowing for further fine-tuning to align with the specifics of the dataset at hand.

To evaluate the potential contribution of external images during training, we conducted experiments wherein the initial 15 epochs out of the last 30 included the external images, while the subsequent 15 epochs exclusively involved the National Museum's dataset. The outcome of this investigation, revealing the optimal configuration, is presented in Table 1. Notably, in most instances, the inclusion of external images did not confer any discernible performance benefits. This outcome is likely attributed to factors such as lower resolution and specific subject composition in the external images.

Table 1 summarises the optimal parameterisation options for each of the DL techniques considered.

VI. RESULTS

Figure 9 provides a visual representation of the performance of the various DL techniques outlined in Table 1. The 0-notch performance accuracy represents the model's accuracy when its prediction corresponds to the class with the highest output probability, a principle commonly known in the machine learning literature as 'winner-takes-all.' The 1-notch and 2-notch accuracies, on the other hand, measure whether the true target lies within the top two or three most likely predictions generated by the DL model. As expected, the plain vanilla CNN model yields the lowest performance across all accuracy metrics closely followed by the ConvMixer model. The ConvNeXt model emerges as the top performer, achieving impressive results with a 0-notch accuracy of 79.62%, a 1-notch accuracy of 94.86%, and a 2-notch accuracy of 97.77%. These results are particularly noteworthy, especially when considering the challenge of classifying images into eight distinct categories.

Table 2 displays the confusion matrix of the trained ConvNeXt model on the test set. Correct classifications appear on the diagonal and are depicted in bold face. Off diagonal elements correspond to misclassification errors. Let's now summarize the confusion matrix in terms of precision and recall. Precision assesses the accuracy of a prediction, while recall gauges its completeness. For instance, we can calculate the precision for cruisers by dividing the correct predictions by the corresponding column sum, which results in 92/(23+2+2+92+18+2+4) = 64.34%. This means that we can have 64.34% confidence that when the model predicts a cruiser, it is correct. The recall for cruisers is calculated by dividing the correct predictions by the corresponding row sum, yielding 92/(5+1+92+2) = 92%. This figure indicates that 92% of the actual cruisers are correctly classified as

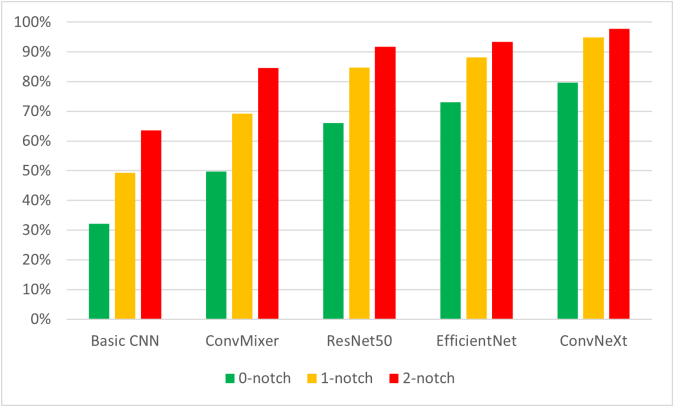


FIGURE 9: 0-, 1-, 2-notch accuracies of DL models.

cruisers.

It's not surprising that submarines and carriers exhibit the highest precision and recall, as they are relatively easy to identify due to their distinctive architectural features. In fact, we discovered that only one submarine was misclassified as a carrier. Further examination revealed that this was the HMS Truculent, as depicted in Figure 10. Its somewhat unconventional design, characterized by an unusual fin and armament, likely led the ConvNeXt model to misclassify it as a carrier. Corvettes, on the other hand, rank among the lowest in terms of both precision and recall, as they are challenging to differentiate from other vessel types, such as frigates.

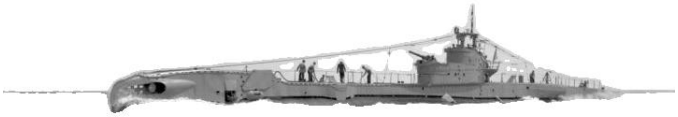


FIGURE 10: Submarine HMS Truculent (1946) erroneously classified as a carrier.

VII. INTERPRETING DEEP LEARNING MODELS

Deep learning models are widely recognised for their inherent complexity and opacity, rendering them challenging to interpret due to the intricate mathematical transformations they employ to map input data (such as ship images) to output results (like ship classification). The integration of explanatory mechanisms into artificial intelligence models, particularly deep learning, falls within the realm of Explainable AI (XAI) [39]–[41]. The incorporation of XAI techniques into image classification models offers a multitude of advantages. Foremost among these benefits is the cultivation

Technique	Transfer Learning	External Data	Focal Loss	Key Parameters
Basic CNN	No	Yes	Yes	2 convolution layers, 2 dense layers, dropout=40%, dense head
ConvNeXt	Yes	No	No	2D global average pooling layer, dropout=40%, dense head
EfficientNet	Yes	No	Yes	2D global average pooling layer, dropout=40%, dense head
ResNet	Yes	No	No	dropout=40%, dense head
ConvMixer	No	Yes	Yes	2D global average pooling layer, dropout=40%, dense head

TABLE 1: Optimal DL method configurations.

		Predicted							
		BS	**CAR**	**CORV**	**CRUIS**	**DEST**	**FRIG**	**MINE**	**SUB**
Actual	**BS**	**54**	0	0	23	0	1	0	0
	CAR	0	**41**	0	2	0	0	0	0
	CORV	0	0	**4**	2	0	3	0	0
	CRUIS	5	1	0	**92**	0	2	0	0
	DEST	0	0	2	18	**134**	26	7	0
	FRIG	2	0	0	2	8	**63**	8	0
	MINE	0	0	0	4	1	1	**45**	0
	SUB	0	1	0	0	0	0	0	**32**

TABLE 2: Confusion matrix contrasting actual versus predicted classes. Note: BS = Battleship, CAR = Carrier, CORV = Corvette, CRUIS = Cruiser; DEST = Destroyer, FRIG = Frigate, MINE = Minesweeper, SUB = submarine.

	Precision	Recall
Battleship	88,52%	69,23%
Carrier	95,35%	95,35%
Corvette	66,67%	44,44%
Cruiser	64,34%	92,00%
Destroyer	93,71%	71,66%
Frigate	65,63%	75,90%
Minesweeper	75,00%	88,24%
Submarine	100,00%	96,97%

TABLE 3: Precision and Recall.

of trust among decision-makers, in this case curators and archivists, within our application. Moreover, it equips them with valuable guidance when verifying classifications, especially in situations where the deep learning model exhibits uncertainty, possibly due to factors like image distortion or low resolution. By drawing upon prior explanations of images, XAI informs decision-makers about the key elements the model focused on, clarifying its decision-making process.

Various techniques have been proposed to shed light on the internal functioning of a DL model trained for image classification. For instance, [42] and [43] introduced model-agnostic methods for understanding image classifications using counterfactual explanations. [44] introduced an uncertainty quantification-based framework to interpret DL decisions for image classification. [45] surveyed more than 200 papers using XAI methods for deep learning-based medical image analysis. In this study, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [46] to explain the ConvNeXt classifications since it yields very intuitive visual explanations of the classifications made which largely contribute its success. Grad-CAM uses the gradients of any target class (e.g., cruiser, submarine, carrier, etc), flowing into the final convolutional feature map to produce a coarse localisation map highlighting the important image regions which are key to predict the target. In other words, large gradients correspond to image segments which highly contribute to the final classification. Besides its attractive visualisation, one of its key benefits is that it's applicable to a wide variety of CNN architectures, such as our ConvNeXt models. Figure 11 displays some examples of Grad-CAM heatmaps for some of the ships in our data set made using Python's grad-cam package.

The Grad-CAM heatmap for HMS E11, a submarine with a pivotal role in the Dardanelles battle (1915-1916), clearly highlights the essential parts of the sub. Notably, the emphasized regions align seamlessly with the submarine's fin and sections of its casing, which are the standard visible areas of a submarine when it is surfaced. Adjacent to this, the heatmap for the HMS Ark Royal aircraft carrier highlights its distinctive ski-jump ramp. The bottom left of Figure 11 showcases the plot for the HMS Gavinton minesweeper, where the conspicuous highlights clearly demarcate the ship's minesweeping equipment, positioned at the stern. On the bottom right, we observe the heatmap for the HMS Dreadnought battleship, presenting a discernible edge detection filter extending up to the ship's top antenna. These visual representations vividly underscore the enhanced explanatory capacity of Grad-CAM heatmaps in demystifying intricate image classifications achieved through deep learning. They offer a visually intuitive and accessible means of comprehension for human decision-makers, in this case curators and archivists.

VIII. CONCLUSION

In this paper, we conducted a comprehensive exploration of deep learning techniques for warship classification, utilising a distinct and exclusively obtained dataset from the National Museum of the Royal Navy. Specifically, our investigation involved the development of an image classification sys-

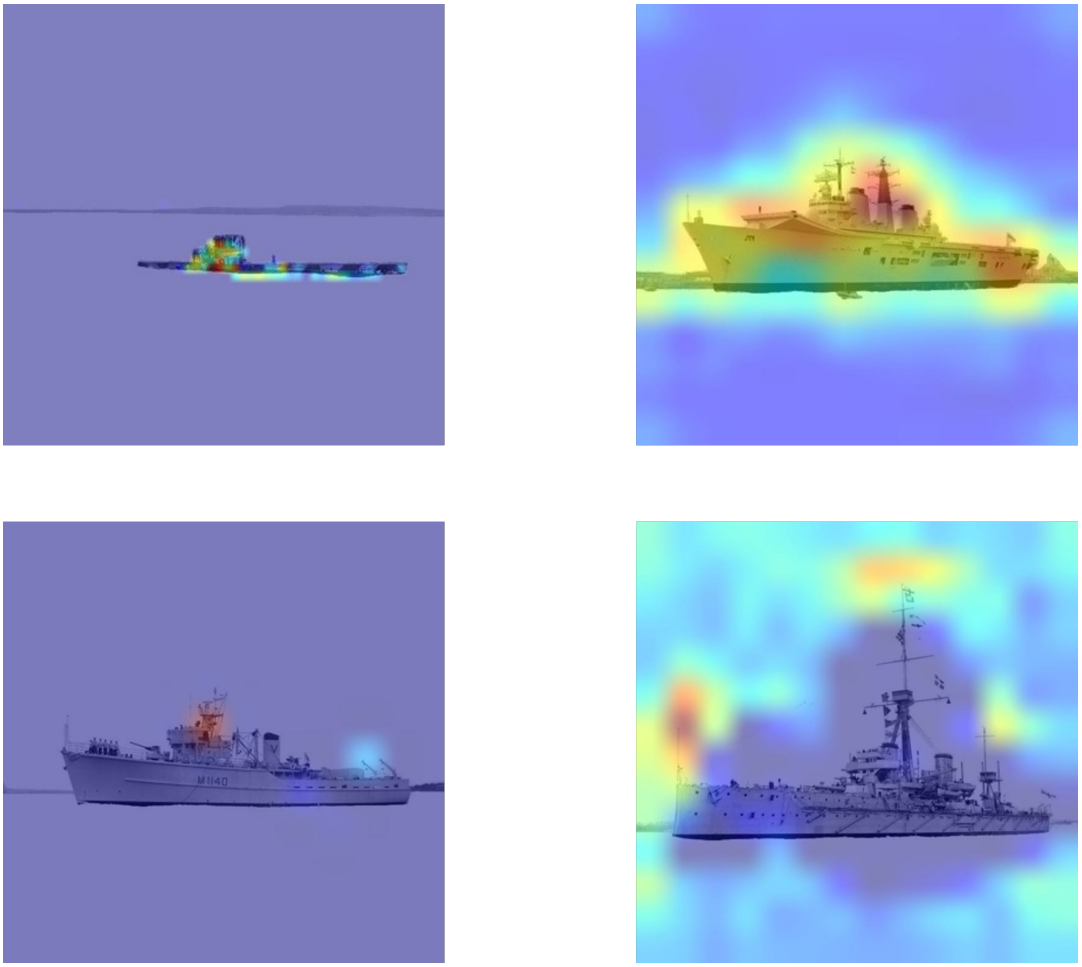


FIGURE 11: Example Grad-CAM images. Top Left: HMS E11 submarine ($\sim$ 1914); Top Right: HMS Ark Royal carrier (1993); Bottom Left: HMS Gavington minesweeper ($\sim$ 1952); Bottom Right: HMS Dreadnought battleship (1906).

tem using pre-trained convolutional neural networks such as ConvNeXt, ResNet, and EfficientNet, juxtaposed with traditional CNNs and ConvMixer models. Our findings, in terms of classification performance, revealed ConvNeXt as the standout performer, achieving an accuracy of 79.62% for 0-notch classification and an impressive 94.86% for 1-notch classification.

We delved into various data preprocessing strategies to enhance performance. Interestingly, the inclusion of externally acquired images did not yield discernible benefits, while image segmentation, by effectively eliminating irrelevant image components, yielded positive effects. Additionally, the application of safe augmentations, such as horizontal image flipping, sharpening, rotation, and resizing, proved to be advantageous. Furthermore, we employed Grad-CAM to demonstrate how ConvNeXt's complex, opaque models could be rendered more interpretable for archivists, curators and documentation managers, offering visually appealing insights which can be nicely deployed into a decision support system.

Our study presents a myriad of avenues for future research. We aim to expand our dataset further through ongoing digitisation efforts at the National Museum of the Royal Navy. Additionally, a more fine-grained classification, encompassing more diverse ship types, such as landing ships, repair ships, amphibious assault ships, sloops, survey vessels, trawlers, and depot ships, is a promising area for exploration. While involving more resources, human labellers may contribute to improved image segmentation. Beyond Grad-CAM, we encourage research into other explanatory methods, such as counterfactuals. Ultimately, we believe that our proposed methodology can readily find applications in various other image classification domains.

REFERENCES

[1] N. Díaz-Rodríguez and G. Pisoni. Accessible cultural heritage through explainable artificial intelligence. In Workshop at the 28th ACM UMAP Conference on User Modeling, Adaptation and Personalization, 05 2020.
[2] J. Liu, X. Kong, F. Xia, X. Bai, L. Wang, Q. Qing, and I. Lee. Artificial intelligence in the 21st century. IEEE Access, 6:34403–34421, 2018.
[3] A.L. Cushing and G. Osti. So how do we balance all of these needs?": how the concept of ai technology impacts digital archival expertise. Journal of Documentation, 79(7):12–29, 2023.
[4] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. Nature, 532(7553):436–444, 2015.

[6] A. Shrestha and A. Mahmood. Review of deep learning algorithms and architectures. IEEE Access, 7:53040–53065, 2019.

[7] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. 01 1995.

[8] J. Plested and T. Gedeon. Deep transfer learning for image classification: a survey, 05 2022.

[9] G. Aracri, A. Folino, and S. Silvestri. Integrated use of kos and deep learning for data set annotation in tourism domain. Journal of Documentation, Jan 2023.

[10] M. Reusens, A. Stevens, J. Tonglet, W. Verbeke, S. vanden Broucke, and B. Baesens. Evaluating text classification: A benchmark study. Submitted for publication, 2023.

[11] J. Ker, L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. IEEE Access, 6:9375–9389, 2018.

[12] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11966–11976, 2022.

[13] A. Trockman and J.Z. Kolter. Patches are all you need?, 2022.

[14] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6105–6114. PMLR, 09–15 Jun 2019.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[20] J. Deng, W. Dong, R. Socher, L.J. Li, L. Kai, and F.F. Li. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[21] D. Guo, D. Gao, Z. Chen, Y. Li, X. Zhao, W. Song, and L. Xiaolei. Classification of inbound and outbound ships using convolutional neural networks. Frontiers in Marine Science, 10, 05 2023.

[22] S. Hashimoto, Y. Sugimoto, K. Hamamoto, and N. Ishihama. Ship classification from sar images based on deep learning. In K. Arai, S. Kapoor, and R. Bhatia, editors, Intelligent Systems and Applications, pages 18–34, Cham, 2019. Springer International Publishing.

[23] B. Li, X. Xie, X. Wei, and W. Tang. Ship detection and classification from optical remote sensing images: A survey. Chinese Journal of Aeronautics, 34(3):145–163, 2021.

[24] A. Hazarika, P. Jidesh, and A. Smitha. Comparative analysis of machine learning and deep learning models for ship classification from satellite images. In B. Raman, S. Murala, A. Chowdhury, A. Dhall, and P. Goyal, editors, Computer Vision and Image Processing, pages 60–72, Cham, 2022. Springer International Publishing.

[25] R. Pacheco-Ruiz, J. Adams, and F. Pedrotti. 4d modelling of low visibility underwater archaeological excavations using multi-source photogrammetry in the bulgarian black sea. Journal of Archaeological Science, 100:120–129, 2018.

[26] R. Pacheco-Ruiz, J. Adams, F. Pedrotti, M. Grant, J. Holmlund, and C. Bailey. Deep sea archaeological survey in the black sea–robotic documentation of 2,500 years of human seafaring. Deep Sea Research Part I: Oceanographic Research Papers, 152:103087, 2019.

[27] J. McCarthy, J. Benjamin, T. Winton, and W. Van Duivenvoorde. The rise of 3d in maritime archaeology. 3D Recording and Interpretation for Maritime Archaeology, pages 1–10, 2019.

[28] T. Brughmans and M A. Peeples. pages i–i. Cambridge Manuals in Archaeology. Cambridge University Press, 2023.

[29] M. Leclerc, R. Tharmarasa, M. Florea, A.C. Boury-Brisset, T. Kirubarajan, and N. Duclos-Hindie. Ship classification using deep learning techniques for maritime target tracking. Information Fusion, pages 750–757, 07 2018.

[30] L. Leonidas and Y. Jie. Ship classification based on improved convolutional neural network architecture for intelligent transport systems. Information, 12:302, 07 2021.

[31] L. Zhenzhen, Z. Baojun, T. Linbo, L. Zhen, and F. Fan. Ship classification based on convolutional neural networks. The Journal of Engineering, 2019, 10 2019.

[32] N. Mishra, A. Kumar, and K. Choudhury. Deep convolutional neural network based ship images classification. Defence Science Journal, 71:200–208, 03 2021.

[33] Z. Xu, J. Sun, and Y. Huo. Ship images detection and classification based on convolutional neural network with multiple feature regions. IET Signal Processing, 16, 02 2022.

[34] Y. Yang, K. Ding, and Z. Chen. Ship classification based on convolutional neural networks. Ships and Offshore Structures, 17(12):2715–2721, 2022.

[35] S. Vanden Broucke and B. Baesens. Practical Web Scraping for Data Science. Apress, 2018.

[36] Janes Information Services. Jane's Fighting Ships. 2023.

[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, P. Dollár, and R. Girshick. Segment anything, 2023.

[38] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2999–3007, 2017.

[39] K. W. De Bock, K. Coussement, A. De Caigny, R. Słowiński, B. Baesens, R. N Boute, T. M. Choi, D. Delen, M. Kraus, S. Lessmann, S. Maldonado, D. Martens, M. Óskarsdóttir, C. Vairetti, W. Verbeke, and R. Weber. Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda. European Journal of Operational Research, September 2023.

[40] E. Carrizosa, J. Ramirez-Ayerbe, and D.R. Morales. Generating collective counterfactual explanations in score-based classification via mathematical optimization. Expert Systems with Applications, 238:121954, 2024.

[41] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access, 6:52138–52160, 2018.

[42] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens. Explainable image classification with evidence counterfactual. Pattern Analysis and Applications, 25(2):315–335, May 2022.

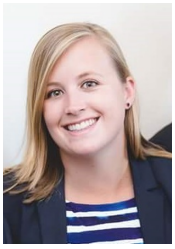
[43] M. Barbosa de Oliveira, R, K. Sörensen, and D. Martens. A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data. European Journal of Operational Research, 2023.

[44] X. Zhang, F.T.S. Chan, and S. Mahadevan. Explainable machine learning in image classification models: An uncertainty quantification perspective. Knowledge-Based Systems, 243:108418, 2022.

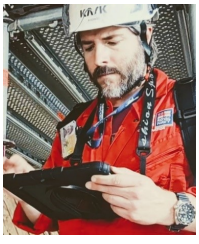
[45] B.H.M. van der Velden, H.J. Kuijf, K.G.A. Gilhuijs, and M.A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis, 79:102470, 2022.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017.

BART BAESENS is a professor of Data Science at KU Leuven (Belgium), and a lecturer at the University of Southampton (United Kingdom). He has done extensive research on data science, credit risk modeling, fraud detection, and marketing analytics. He co-authored more than 250 scientific papers and 10 books. Bart received the OR Society's Goodeve medal for best JORS paper in 2016 and the EURO 2014 and EURO 2017 award for best EJOR paper. His research is summarized at www.dataminingapps.com. Bart is listed in the top 2% of Stanford University's new Database of Top Scientists in the World. He was also named one of the World's top educators in Data Science by CDO magazine in 2021 and 2023, and has educated tens of thousands of data scientists across the globe. Bart also has his own ON-LINE learning BlueCourses platform: www.bluecourses.com which features courses on machine learning, credit risk, fraud, marketing, text analytics, deep learning, web scraping etc.



AMY ADAMS is the Collections Information & Access Manager at the National Museum with extensive experience in collections management including digital collections. She is a qualified museum professional with a M.Sc from the University of Leicester and a trained historian with a M.A in History from Western University. In her decade with the National Museum, she has been focused on leading and developing its archival collections. She has also spearheaded many digital collection projects including unify collections databases, introduced new digitisation methods, digital asset management and preservation practices. More recently she has been spearheading a number of projects trialling AI technologies in Collections Management practices. She has spoken of her work at various conferences including MuseumNext, Archives Records Association, Museum Computer Group, Museums + Heritage show, among others.



RODRIGO PACHECO-RUIZ is currently the Archaeological Data Manager for HMS Victory at The National Museum of The Royal Navy in Portsmouth UK, where he is in-charge of the archaeological documentation of HMS Victory's Conservation Management Plan and the Victory Archives. This is a long-term project with the aim of restoring the only surviving example of a first-rate ship of the line to her appearance in 1805 using archaeology specific UAS robotic data, high-resolution laser scan datasets and traditional shipbuilding techniques. He is a Visiting Fellow in Maritime Archaeology at the University of Southampton and Co-Director of the Offshore Archaeological Research Progamme (OAR). In 2019 he led the discovery of one of the best-preserved 16th Century shipwrecks in the world (Okänt Skepp) and in 2020 one of the oldest, and long lost, German WWI U-boat (UC-47) offshore of the UK. Rodrigo is a specialist in deep sea archaeology and underwater digital archaeological recording through the use of state-of-the-art robotics. He is also a member of the National Oceanography Centre of the University of Southampton as well as co-investigator at the Mexico's Universidad Nacional Autonoma de Mexico (UNAM) Maritime Archaeology programme as well as an Associate Fellow of the Maritime Archaeology Research Institute (MARIS) of Södertorn University, Sweden. Rodrigo is an HSE Surface Supplied Commercial Diver a Civil Aviation Association Commercial drone pilot and a Nautical Archaeology Society tutor.

ANN-SOPHIE BAESENS is a bachelor student at KU Leuven. She is passioned by new technological developments (e.g., Artificial Intelligence, Deep Learning) as well as British naval history.

SEPPE VANDEN BROUCKE is currently working as an assistant professor at the Department of Business Informatics and Operations Management at UGent (Belgium) and is a guest lecturer at KU Leuven (Belgium). Seppe's research interests include business data mining and analytics, machine learning, process management, process mining. His work has been published in wellknown international journals and presented at top conferences. Seppe received his PhD in Applied Economics at KU Leuven, Belgium in 2014.