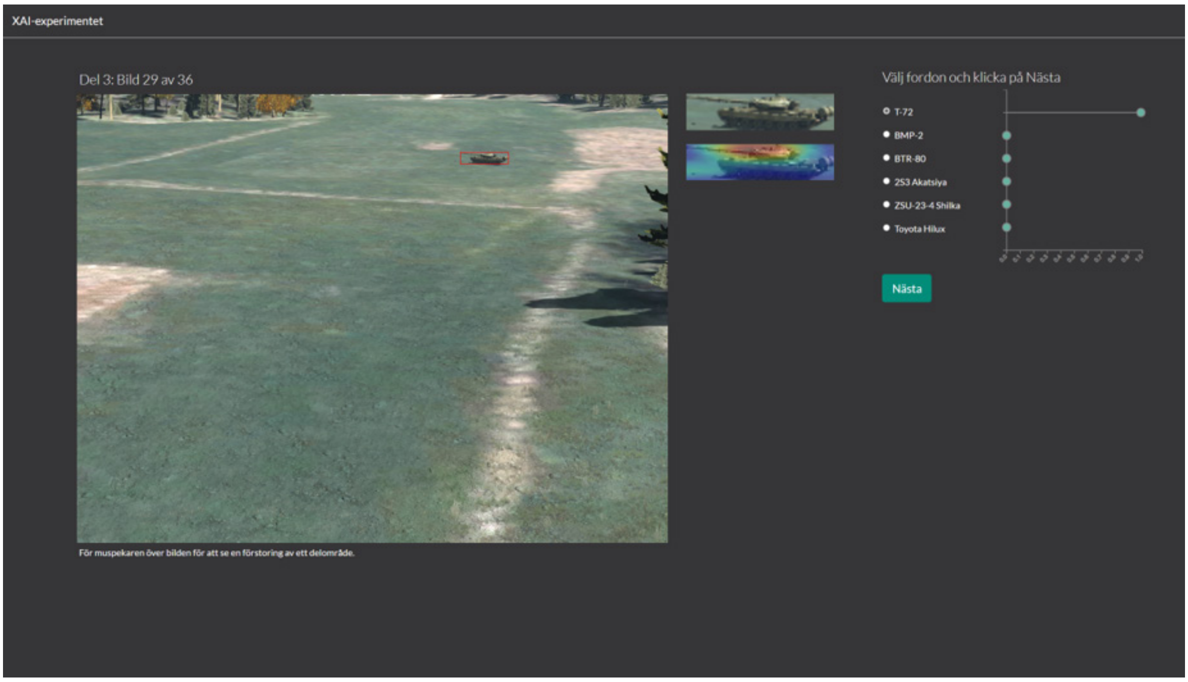
Evaluation of Target Classification Performance of UAV Imagery with RISE Saliency Map Explanations

PETER SVENMARCK, ULRIKA WICKENBERG BOLIN,
DANIEL OSKARSSON, ROGIER WOLTJER



XAI-experimentet

Del 3: Bild 29 av 36

Välj fordon och klicka på Nästa

För muspekaren över bilden för att se en förstoring av ett delområde.

FOI-R--5624--SE
ISSN 1650-1942 Augusti 2024

Peter Svenmarck, Ulrika Wickenberg Bolin,
Daniel Oskarsson, Rogier Woltjer

Evaluation of Target Classification Performance of UAV Imagery with RISE Saliency Map Explanations

Title	Evaluation of Target Classification Performance of UAV Imagery with RISE Saliency Map Explanations
Titel	Utvärdering av prestation för måligenkänning av UAV-bilder med RISE särdragsförklaringar
Rapportnr/Report no	FOI-R--5624--SE
Månad/Month	Augusti
Utgivningsår/Year	2024
Antal sidor/Pages	34
ISSN	1650-1942
Uppdragsgivare/Client	Försvarsmakten
Forskningsområde	Ledningsteknologi
FoT-område	Ledning och MSI
Projektnr/Project no	E38519
Godkänd av/Approved by	Emil Hjalmarson
Ansvarig avdelning	Cyberförsvar och ledningsteknik

Bild/Cover: Screenshot from the web interface used in the experiments in this report.

Summary

Target classification of imagery from Unmanned Aerial Vehicles (UAVs) is increasingly important for military reconnaissance and surveillance. A promising technique to improve target classification of UAV imagery is Deep Neural Networks (DNNs). However, DNNs may consist of a very large number of parameters, which makes it difficult for operators to understand what image features DNNs use for target classification. This lack of transparency is a challenge for military applications of DNNs for target classification since operators are ultimately responsible for all decisions due to the high risks of weapon engagements. Operators therefore also need explanations of DNN classifications to assess their reliability.

This report describes an experiment where participants performed a target classification task of military vehicles in low-altitude UAV imagery. The objective of the experiment was to evaluate whether support of DNN classifications and support of saliency map explanations of DNN classifications, which highlight the most important features for DNN classifications, improve accuracy in target classifications. Saliency map explanations were generated with the Randomized Input Sampling for Explanation (RISE) method. Participants performed the target classification task in three different conditions: without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations of the DNN classifications.

The results show that, contrary to expectations, participants' accuracy in target classification decreases with support of DNN classifications and it decreases even further with support of RISE saliency map explanations. Participants' lower accuracy in target classification with support of DNN classifications and RISE saliency map explanations is likely due to a combination of two reasons: reliance on automated decision aids and difficulty in assessing DNN reliability. The results show that participants under-rely on DNN classifications when they are correct and over-rely on DNN classifications when they are incorrect.

The conclusion of the experiment is that it is not trivial to present DNN classifications and explanations of DNN classifications that actually support operators' target classification. Additional experiments are required of how to present information from DNN classifications and whether other promising XAI-approaches improve operators' target classification.

Keywords: artificial intelligence, deep learning, deep neural networks, XAI, saliency map explanations, RISE, target classification, unmanned aerial vehicles

Sammanfattning

Måligenkänning av bilder från obemannade flygfarkoster (eng. Unmanned Aerial Vehicle, UAV) blir allt viktigare för militär spaning och underrättelseinhämtning. En lovande teknik för att förbättra måligenkänning av UAV-bilder är djupa neuronnät. Problemet är att djupa neuronnät består ett stort antal parametrar som gör det svårt för operatörer att förstå vilka särdrag i bilden som neuronnätet använder för måligenkänning. Bristen på transparens är en utmaning för militära tillämpningar av måligenkänning med djupa neuronnät eftersom operatörer i slutändan är ansvariga för alla beslut på grund av de stora riskerna med väpnade insatser. Operatörer behöver därför även förklaringar av neuronnätets måligenkänning för att bedöma dess tillförlitlighet.

Rapporten beskriver ett experiment där deltagarna genomförde måligenkänning av militära fordon i UAV-bilder tagna på låg höjd. Syftet med experimentet var att utvärdera om stöd av ett djup neuronnät och stöd av särdragsförklaringar (eng. saliency maps), som markerar framträdande särdrag för djupa neuronnäts måligenkänning, förbättrar måligenkänningen. Särdragsförklaringarna skapades med metoden *Randomized Input Sampling for Explanation* (RISE). Deltagarna genomförde måligenkänningen under tre förutsättningar: utan stöd av måligenkänningar från ett djupt neuronnät, med stöd av måligenkänningar från ett djupt neuronnät och med stöd av RISE särdragsförklaringar av det djupa neuronnätets måligenkänningar.

Resultaten visar att tvärtemot förväntningarna så minskar deltagarnas förmåga att korrekt känna igen mål med stöd av måligenkänningar från ett djupt neuronnät och den minskar ytterligare med stöd av RISE särdragsförklaringar. Försämringen av deltagarnas måligenkänning beror sannolikt på en kombination av två orsaker: förlitan till automatiserade beslutsstöd och svårighet att bedöma neuronnätets tillförlitlighet. Resultaten visar att deltagarna har för låg förlitan till korrekta klassificeringar från neuronnätet och för hög förlitan till inkorrekta klassificeringar från neuronnätet.

Slutsatsen från experimentet är att det inte är trivialt att presentera måligenkänningar från ett djupt neuronnät och förklaringar av neuronnätets måligenkänningar som faktiskt förbättrar operatörers måligenkänning. Ytterligare experiment behövs av hur information från neuronnätets måligenkänning ska presenteras och om andra lovande XAI-metoder förbättrar operatörers måligenkänning.

Nyckelord: artificiell intelligens, djupinlärning, djupa neuronnät, XAI, särdragsförklaringar, RISE, måligenkänning, obemannade flygfarkoster

Contents

1 Introduction

Target classification in imagery from low-altitude Unmanned Aerial Vehicles (UAVs) is increasingly important for military reconnaissance and surveillance (Rubio, 2020; Kullab, 2023). Target classification in UAV imagery is, however, a challenging task where targets may appear at many locations within the field of view, at large distances, and in varying orientations. Other factors that affect target classification are target similarity, whether the sensors utilize visual or infrared light, as well as other factors, such as time pressure (Lif et al., 2018; Lif et al., 2021).

A promising technique to improve target classification in UAV imagery is Deep Neural Networks (DNNs) (Mittal et al., 2020; Wu et al., 2021). However, since DNNs may consist of a very large number of parameters, it is often difficult for operators to understand which specific image features are most important for the DNNs' target classification. This lack of transparency is a challenge for military applications where operators are ultimately responsible for all decisions due to the high risks of weapon engagements (Svenmarck et al., 2018). Operators therefore also need explanations of DNN classifications to be able to assess their accuracy.

The goal of eXplainable Artificial Intelligence (XAI) is to make models, such as DNNs, more comprehensible and transparent to humans (Luotsinen et al., 2019; Ali et al., 2023). XAI is important in sensitive applications, such as military, finance, and healthcare, where domain specialists ultimately are responsible for any decisions using the proposed classifications. XAI enables domain specialists to understand and trust models, as well as to assess their accuracy (Hoffman et al., 2019).

A common XAI approach to increase the transparency of DNNs for classification of images is saliency map explanations, which highlight the pixels that are most important for the DNN's classification. Several methods have been proposed for calculation of such saliency map explanations (Vilone & Longo, 2020; Das & Rad, 2020; Luotsinen et al., 2019). Most saliency map methods are intended for DNNs that classify whether an object of interest is present in an image. Saliency map methods for DNNs that classify every object of interest in an image, which is the actual task that operators perform, have only been proposed recently (Tsunakawa et al., 2019; Petsiuk et al., 2021). Further, saliency map methods should preferably be applicable to the wide range of DNNs that are used for image processing. Saliency map methods should therefore preferably be model-agnostic rather than model-specific, which limits their applicability to only some DNNs.

User evaluations of saliency map explanations show that they can provide many benefits. For example, saliency map explanations increase users' understanding of model predictions in proxy tasks where many examples from training data are presented to inform users about model output (Alqaraawi et al., 2020; Colin et al., 2022). However, although saliency map explanations are intuitive for users, they may not present all information that users need to understand model output. For example, evaluations show that saliency map explanations highlight where users should look, but not what finer details within highlighted regions they should look at (Rudin, 2019; Gahsemmi et al., 2021; Colin, 2022). Saliency map explanations are therefore not distinct enough for users to detect incorrect predictions (Kim et al., 2022). Additionally, saliency map explanations only enable inconclusive rejection of spurious models that react to features without any meaningful connection to the actual task (Adebayo et al., 2022). The mixed results about pros and cons of saliency map explanations means that it is unclear if they reduce users' tendency to trust and rely on automated decision aids (Dzindolet et al., 2003) or exacerbate the over-reliance on model predictions (Alqaraawi et al., 2020; Kim et al., 2022). Commonly, such negative effects of XAI approaches are referred to as XAI pitfalls (Ehsan & Riedl, 2021).

With the increasing use of DNNs for target classification it becomes important to evaluate how DNNs by themselves, as well as in combination with saliency map explanations, may improve performance of operator target classification. We therefore report an experiment where participants performed a target classification task of military vehicles in images to evaluate whether DNN classifications and saliency map explanations of the DNN classifications improve target classification performance. The experiment focused on saliency map explanations of a DNN that classify whether an object is present in an image since these saliency map methods are more mature than methods for DNNs that classify every object of interest in an image.

1.1 Experiment objective and hypotheses

The objective of the experiment was to evaluate whether support of DNN classifications and support of saliency map explanations of DNN classifications improve accuracy in target classification of military vehicles in low-altitude UAV imagery. Saliency map explanations were generated with the Randomized Input Sampling for Explanation (RISE) method (Petsiuk et al., 2018). RISE is a model-agnostic saliency map method that has been found to perform fairly well on most metrics for saliency map explanations, such as finding discriminative features, finding features that are most important for classification, discrimination between classes, and small effects of insignificant variations (Li et al., 2021). Participants performed the target classification task in three different conditions: without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations of the DNN classifications. Participants also performed the target classification task with two levels of vehicle resolution. Higher vehicle resolution typically increase the number of features for target classification, although the effect varies depending on the vehicle class.

Since saliency map explanations may be used to detect when DNN classifications are incorrect, one-third of the images were selected so that the DNN classification was incorrect. Without this image selection, the accuracy of the DNN classifications were simply too high for any meaningful number of such incorrect target classifications within the constraints of the experiment.

The hypotheses of the experiment was that participants were expected to:

1) Have higher accuracy in target classification with higher vehicle resolution.
2) Have higher accuracy in target classification with support of DNN classifications than without support of DNN classifications.
3) Have higher accuracy in target classification with support of RISE saliency map explanations of DNN classifications than without support of RISE saliency map explanations.
4) Have higher accuracy in target classification with support of RISE saliency map explanations when DNN classifications are incorrect.

1.2 Outline and reading instructions

Chapter 2 describes the method for the experiment in terms of participants, as well as the images of vehicles, DNN classifications, and saliency map explanations that were used in the experiment. Chapter 3 describes the results of the experiment in terms of how investigated factors affect target classification performance, response time, and reliance on DNN classifications. Chapter 4 summarizes the results and describes possible reasons for the achieved target classification performance. Chapter 5 describes the conclusions of the experiment and provides some recommendations of XAI-approaches for image classification to evaluate in future experiments.

Readers who are mainly interested in the target classification task are referred to section 2.3. Readers who are mainly interested in an overview of the results are referred to Figure 4 to Figure 9, which shows the most important results with additional details in the text.

2 Method

The experiment required participants to perform target classification of images from Virtual Battlespace 3 (VBS3). VBS3 is an interactive virtual military training environment developed by Bohemian Interactive Simulations[1]. VBS3 models military and civilian vehicles, weapons, and characters. The images were typical for low-altitude UAV reconnaissance and surveillance.

2.1 Participants

Sixteen participants (11 males and 5 females; mean age 25.8 years and standard deviation 6.1 years) took part in the experiment. None of the participants had any prior experience in classification of military vehicles. Most participants (twelve of sixteen) regularly played computer games that require skills related to target classification, for example, military simulators, real-time simulators, role-playing games, strategy games, first-person shooter games or similar games. Four participants played such computer games more than five hours per week, three 1–2 hours per week, and five 1–3 hours per month. Four participants did not regularly play any such computer games. All participants had adequate vision with or without correction.

2.2 Materials

The materials for the experiment consisted of a PC workstation, DNN classifications, and RISE saliency map explanations of the DNN classifications.

2.2.1 Apparatus

The target classification task was presented on a Philips Brilliance 272B8QJEB/00 27” LCD monitor with a resolution of 2560 × 1440 pixels and a frame rate of 60 Hz. Images were presented using a web based user interface hosted on a custom PC workstation with an Intel Core i7-7800X 3.5 GHz processor and a NVIDIA GeForce RTX 2080 8GB GPU. Images from VBS3, DNN classifications, DNN confidence values, and RISE saliency map explanations were generated beforehand and presented using the web based user interface. The experiment was conducted in a quiet room with dimmed lights.

2.2.2 Dataset, DNN model architecture, and training

The dataset for target classification consisted of images from VBS3 with a resolution 800 × 600 pixels. The images consisted of six vehicle classes: T-72, BMP-2, BTR-80, 2S3, ZSU-23, and Toyota Hilux. Toyota Hilux had a distinct color and shape compared to the other vehicle classes and was included as control classifications to verify that participants focused on the target classification task. The images were sampled from random positions within 12.6 km^2 of a training area for Swedish Armed Forces in VBS3. The vehicles were sampled from random orientations, aspect angles, elevations angles 5º–15º, and 39–143 vehicle resolution in pixels. Varying direction of the sensor relative the vehicle was used for changing vehicle placement within the field of view. Accuracy in vehicle resolution was achieved by sampling vehicle resolution at fixed distances and interpolating a quadratic function. The distance for desired resolution was then calculated using a root finding algorithm. Vehicles occupied about 1 degree of visual angle on the screen. 1,000 images were generated for each vehicle class, totally 6,000 images. 5,400 images were used for training of the DNN classifier and 600 images were used for testing of the classifier. Bounding boxes were generated from highlighting the vehicle shape in VBS3.

[1] https://bisimulations.com/products/vbs3

The DNN classifier was implemented with a convolutional neural network (CNN). The CNN consisted of two convolutional blocks followed by a fully connected (FC) block terminating in softmax activation. Each convolutional block consisted of a convolutional layer (2D) with ReLU activation, a max pooling layer (2D), and finally 30% dropout. The convolutional layers used 64 and 32 2-by-2 filter kernels, respectively. The same kernel size was used for max pooling, but with a stride of 2. The FC block consisted of two FC layers, separated by 50% dropout. The first layer had an input size of 100 × 100 pixels, output size of 256, and ReLU activation. Image regions within ground truth bounding boxes were resized by interpolation to the input size without padding or preservation of aspect ratio.

The CNN was trained on all of the images in the training set for 20 epochs. Training was performed in PyTorch using cross entropy loss and the Adam optimizer. The DNN achieved a mean accuracy of 95% (confidence threshold = .5). The accuracy per class was 96% for T-72, 89% for BMP-2, 89% for BTR-80, 98% for 2S3, 99% for ZSU-23, and 100% for Toyota Hilux.

It should be noted that the CNN architecture described above can be considered shallow in comparison to widely used image classifier architectures with higher performance (e.g. He et al., 2016). However, the DNN classifier's accuracy was still sufficient for the experiment since it was much higher than the participants' accuracy. The infrequent mistakes were also frequent enough to enable evaluation of how RISE saliency map explanations may increase detection of incorrect classifications.

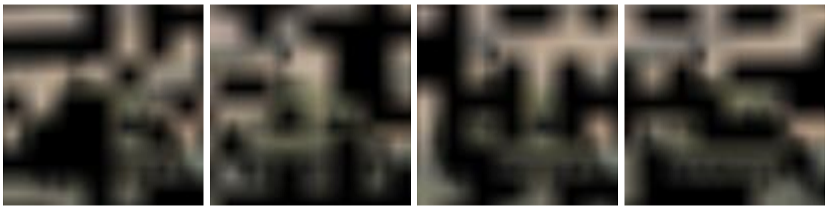
2.2.3 RISE saliency map explanations

RISE is a model-agnostic saliency map method to explain the output predictions of a classifier (Petsiuk et al., 2018). RISE presents explanations in a saliency map that highlights the most important features of the input data that contribute to the classifier's prediction. RISE works by generating a large number of random masks, each of which obscures different subsets of the input pixels, and observing the effect on the classifier's output. Since the classifier's output is more sensitive to obscuring of important pixels, RISE can calculate the importance of input pixels for the classifier's output.

RISE generates masks by upsampling of smaller random binary masks to the input size using bilinear interpolation to create smooth masks without sharp edges when overlaid on the input image. The generation of random binary masks depends on three parameters: the number of binary masks (*N*), the size of the binary masks (*s*) (resolution *s* × *s* pixels), and the probability of a pixel being on (white) in each binary mask (*p*). The output from RISE has been found to be sensitive to these parameters (Stanchi et al., 2023). RISE then upsamples the binary masks to the DNN input size and overlay them on the input image. Figure 1 shows some examples of upsampled binary masks for images with a resolution of 100 × 100 pixels.

Figure 1

Examples of masks with a resolution of 100 × 100 pixels. The masks were generated with the parameters s = 8 and p = .5.

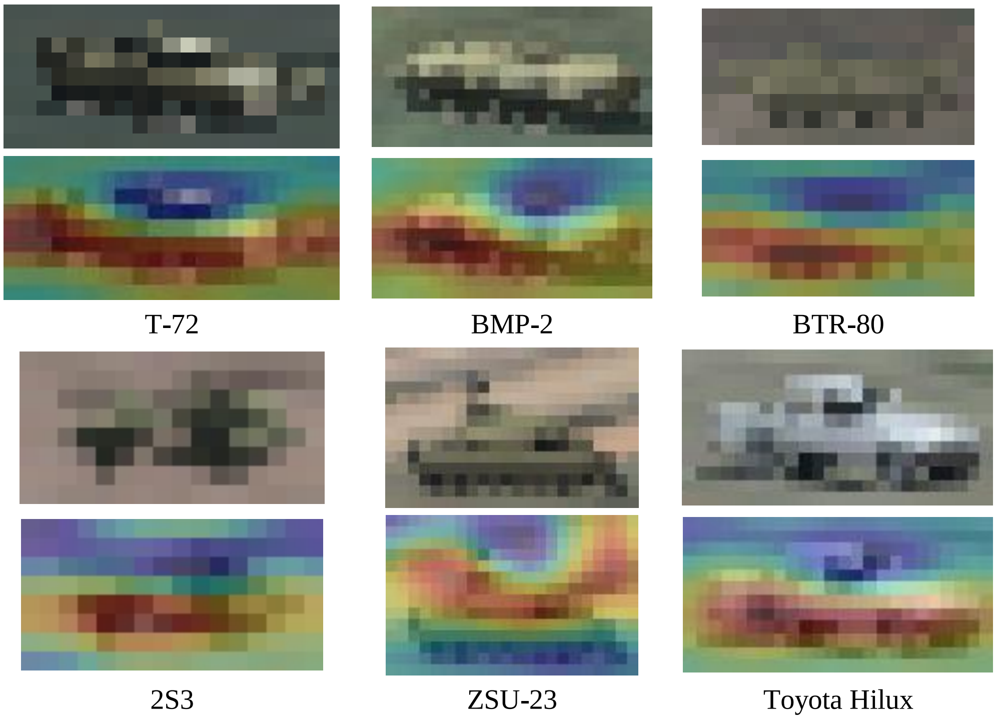


Since suitable parameters depend on the specific classification task, a parameter search was performed for the parameter spans: *N* = [2000, 3000, 4000, 5000, 6000, 7000, 8000], *s* = [4, 6, 8, 10, 12], and *p* = [.1, .2, .3, .4, .5] on 18 images from each vehicle class, 108 images in total. The output results of the model were evaluated using the deletion and insertion tests described by Petsiuk et al. (2018) and further detailed in Luotsinen et al.

(2019). In a deletion test, the most salient features (pixels) are systematically removed to evaluate how the model's performance changes. Conversely, an insertion test involves adding salient features to observe their impact on the model's predictions. The Area Under the Curve (AUC) metric was used to quantitatively compare among results across various settings (Bradley, 1997). For deletion, smaller AUC values are better than larger values. Similarly, for insertion, larger AUC values are better than smaller values. An average of the AUC values was calculated and plotted for each set of parameters (*N*, *s*, and *p*). The AUC values were generated on the 108 images with one parameter fixed and varying values for the other two parameters. The best parameters were found to be $N = 6000$, $s = 12$, and $p = .3$. The calculated RISE saliency map explanations are resized from the DNN input size to the vehicle image size before presentation. Figure 2 shows some examples of RISE saliency map explanations overlaid on images for the six vehicle classes. Red pixels in the RISE saliency map explanations indicate features that were most important for the DNN classification and blue pixels indicate features that were least important for the DNN classification.

Figure 2

Examples of vehicle classes (top rows) and RISE saliency map explanations of the DNN classifications of the vehicle classes (bottom rows).

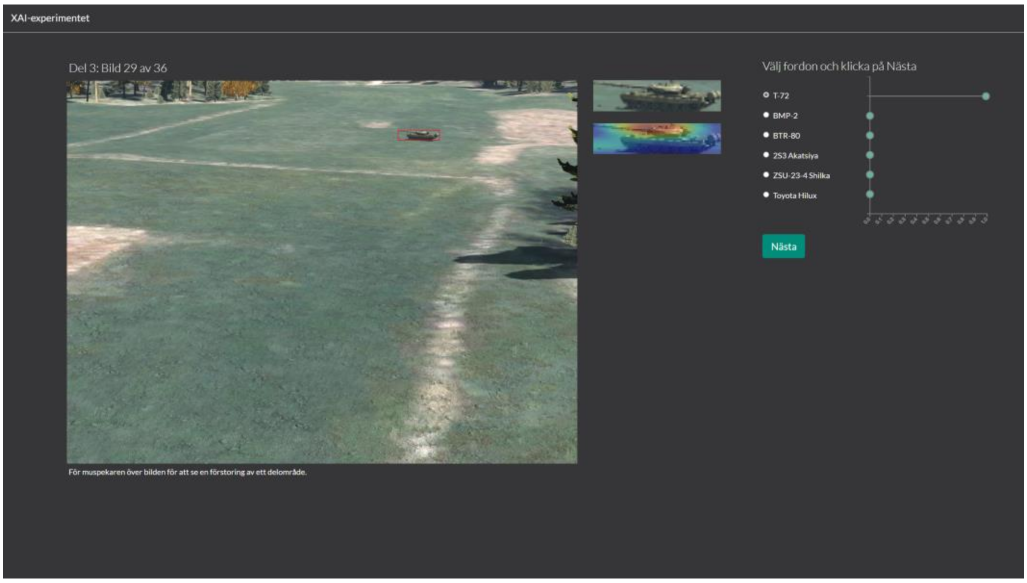


2.3 Target classification task

Participants viewed images of vehicles via a web based user interface for the three conditions: without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations of the DNN classifications. Figure 3 shows the web based user interface for the target classification task with support of a DNN for target classification and a RISE saliency map explanation. The web based user interface for the target classification task with support of a DNN for target classification was similar, but without a RISE saliency map explanation. The web based user interface for the target classification task without support of a DNN for target classification was also similar, but without the bar chart of DNN confidence values and RISE saliency map explanation. Participants used the computer mouse to select from the menu which vehicle class they had identified. There were no limits on response time. A blank screen was presented for five seconds between target classifications to neutralize the effect of the visual stimuli on participants (Bachman & Francis, 2014). The target classification task was similar to Lif et al. (2021) in terms of classification of military vehicles in varying resolutions.

Figure 3

Web based user interface *for the target classification task with support of RISE saliency map explanations of the DNN classifications. Positioned prominently on the left-hand side, a sizable image depicts the vehicle within its environment. A bounding box serves as a visual cue, guiding participants to the target location within the image. To aid in the identification process, a magnifying glass feature activates upon hovering the mouse cursor over the image, particularly useful for discerning vehicles occupying a small visual angle. Adjacent to the main image, an enlarged view of the vehicle, accompanied by its corresponding RISE saliency map explanation, provides further clarity. On the right-hand side, alongside the menu of vehicle classes, a bar chart presents the confidence values of the DNN classification. Each bar represents the probability of the vehicle belonging to a specific class, with values ranging from 0.0 (indicating low probability) to 1.0 (indicating high probability).*



XAI-experimentet

Del 3: Bild 29 av 36

Välj fordon och klicka på Nästa

För muspekaren över bilden för att se en förstoring av ett delområde.

The images for the target classification task were selected from the test set. The images in the test set were only used to verify the classification accuracy of the DNN classifier and thus not previously seen by the DNN classifier during training. Since vehicle classification is notoriously difficult for people in forward and rear facing images of vehicles, only images with favorable aspect angles for vehicle classification were included that showed either the left or right side of vehicles. The images were divided into two vehicle resolution categories, low resolution, which varied between 39 and 72 pixels and medium resolution, which varied between 77 and 143 pixels. Three images were presented for each of the six vehicle classes in low and medium resolution. Participants viewed 36 images in each of the three conditions: without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations of the DNN classifications. In total, each participant viewed 108 images.

Images for the vehicle classes T-72, BMP-2, BTR-80, 2S3, and ZSU-23 were selected to have one-third incorrect DNN classification per vehicle class and resolution. Both correctly and incorrectly classified images were selected based on the criterion of having been difficult for the DNN classifier. An image was considered difficult for the DNN classifier if the DNN confidence values were low for the most likely vehicle class when the vehicle classification was correct and if the DNN confidence values were high for the most likely vehicle class when the vehicle classification was incorrect. Images for the vehicle class Toyota Hilux were selected so that the DNN classification was correct for all images due to a very high accuracy in DNN classification of Toyota Hilux. The selection of incorrect DNN classifications for one-third of the images for five of six vehicle classes meant that the accuracy of the DNN classifier within the experiment was 72.2% and not 66.7%.

2.4 Design of the experiment and statistical analysis

The design of the experiment was intended for evaluation of how participants' performance in target classification was affected by levels of support, vehicle classes, DNN correctness, and vehicle resolution. The design of the experiment was three levels of support (without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations of the DNN classifications), by six vehicle classes, by two DNN correctness (correct vs. incorrect DNN classifications), by two vehicle resolutions (low vs. medium). The design of the experiment was a within subject design, which meant that each participant performed all conditions.

Participants' performance in target classification was analyzed with four-way repeated measure analysis of variance (ANOVA) with factorial design (Field, 2024). The factors in the experiment were levels of support, vehicle classes, DNN correctness, and vehicle resolution. The analysis was performed with repeated measures ANOVA, since each participant performed all conditions. ANOVA uses the variance in terms of how much measures of performance deviate from means to estimate whether the differences between means are statistically significant. The size of the statistical effect is expressed as the *F*-value combined with two numbers for the degrees of freedom that are based on the number of means (*df1*) and number of measures of performance (*df2*) that are included in the analysis. The *F*-value is reported as *F* followed by the value of *df1* and *df2* within parentheses and then the *F*-value. For example, *F*(3, 10) = 5.23. A higher *F*-value means a larger statistical effect. ANOVA also calculates the *p*-value for the probability that the statistical effect did not occur by chance. The criterion for a significant effect was $p < .05$ and the criterion for a statistically significant tendency of effect was $p < .10$. The *p*-value is reported after the *F*-value. For example, *F*(3, 10) = 5.23, $p = .03$. Statistical effects for each factor are referred to as main effects and the combined effects between factors are referred to as interaction effects. An interaction effect typically means that means of the conditions for one factor shows different patterns depending on the other factor or factors.

The t-test was used to test which differences between means were statistically significant (Field, 2024). The t-test is similar to ANOVA, but each t-test only tests the difference between two means. A two-tailed t-test was used since it was not known beforehand which means would be larger than the other means. The *p*-values were adjusted using Bonferroni correction to compensate for the increasing likelihood of detecting significant differences when testing multiple hypotheses. Bonferroni correction adjusts the *p*-value by multiplying it with the number of hypotheses. The same criterions as for ANOVA were used to test for statistically significant differences with the adjusted $p < .05$ for a statistically significant effect and adjusted $p < .10$ for a statistically significant tendency of effect.

2.5 Dependent variables

The dependent variables are the measures that were used to measure participants' performance in target classifications. The dependent variables were accuracy in target classification, response time, reliance on DNN classification, and questionnaires for mental workload, trust, and satisfaction of explanations.

2.5.1 Accuracy in target classification

Accuracy in target classification was measured as the percentage of correctly classified vehicles relative to the total number of vehicles.

2.5.2 Response time

Response time for target classification was measured as the time in seconds for target classification.

2.5.3 Reliance on DNN classification

Reliance on the DNN classification was measured as the percentage of target classifications that were identical to the DNN classification. Reliance on the DNN classification is a misnomer for the condition without support of DNN classifications since classifications were not shown to participants in this condition. However, the percentage of identical target classifications still served as a baseline for comparison with the other two conditions where DNN classifications were shown to the participants.

2.5.4 Questionnaires

Four questionnaires were used to measure participants' subjective experience of the support of DNN classifications and RISE saliency map explanations (see Appendix). Mental workload was measured with an adapted version of NASA-TLX (Hart & Staveland, 1988), compromised of six items on a ten-point scale (1–10). Trust in DNN classifications and RISE saliency map explanations were both measured with six items from the Trust Scale recommended for XAI (adapted from Hoffman et al., 2019). Satisfaction of RISE saliency map explanations was measured with five items from the Explanation Satisfaction Scale (adapted from Hoffman et al., 2019). The trust and satisfaction scales were compromised of a seven-point Likert scale (1–7) from strongly disagree to strongly agree.

2.6 Procedure

Prior to the experiment, participants completed an informed consent form and a general background form. Participants then completed several steps as preparation for the target classification task. In the first step, participants were instructed about distinguishing features between vehicle classes, the target classification task, and the web based user interface, without any support of neither DNN classifications nor RISE saliency map explanations. For the participants, the DNN classifier was referred to as *The Classifier* and RISE saliency map explanations as *Heatmap explanations*. In the second step, participants practiced on 36 target classifications, six target classifications for each vehicle class, without support of DNN classifications. In the third step, participants were instructed about the web based user interface with support of DNN classifications and RISE saliency map explanations. In the fourth step, participants practiced on six target classifications, one target classification for each vehicle class, with support of DNN classifications. In the final step, participants practiced on six target classifications, one target classification for each vehicle class, with support of RISE saliency map explanations of the DNN classifications.

Participants then performed the target classification task. The presentation order of images was randomized over vehicle classes and resolutions, while the order of conditions in terms of support of DNN classifications and RISE saliency map explanations was balanced between participants. The participants' overall performance in target classification was therefore not affected by the presentation order of images or type of support. The questionnaire for mental workload was administered after every condition. The questionnaire for trust in DNN classifications was administered after each condition where DNN classifications were presented either without or with support of RISE saliency map explanations. The questionnaires for trust in RISE saliency map explanations and satisfaction of RISE saliency map explanations were only administered after the condition where RISE saliency map explanations were presented.

2.7 Imputation of missing values

The design of the experiment was incomplete since the DNN classification of Toyota Hilux was always correct. Consequently, data values for target classification and response time were missing for incorrect DNN classification of Toyota Hilux. The missing data values were imputed with the participants’ mean accuracy and mean response time, respectively, for the correct DNN classification of Toyota Hilux. Separate mean accuracy and mean response time were imputed for each combination of vehicle resolution and type of support.

3 Results

Participants' accuracy in target classification, response time, reliance on DNN classification, and subjective ratings of mental workload, trust, and satisfaction were analyzed with separate repeated measures analysis of variance (ANOVA) with factorial design. This chapter presents the results of these ANOVA.

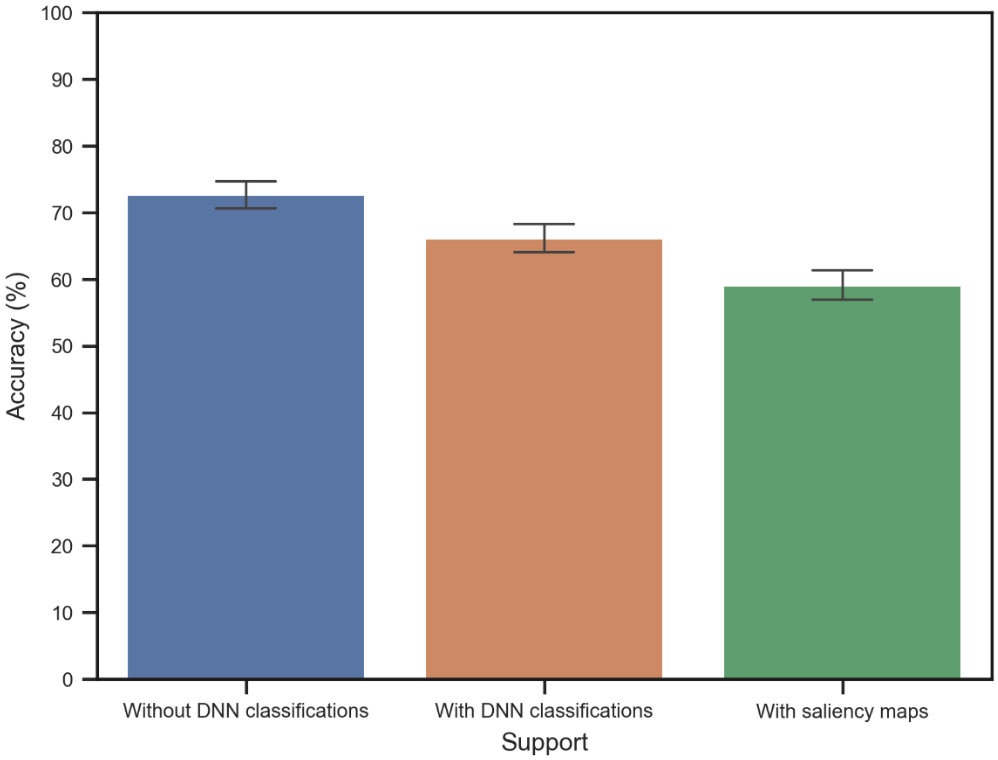
3.1 Accuracy in target classification

Accuracy in target classification was analyzed with a repeated measures 3×6×2×2 ANOVA of the factors support (without DNN classifications, with DNN classifications, and with RISE saliency map explanations), vehicle class (six classes), DNN classification correctness (correct vs. incorrect), and vehicle resolution (low vs. medium).

The analysis of accuracy in target classification showed significant main effects of all four factors. There was a significant main effect of support *F*(2, 39) = 14.6, *p* < .001. However, contrary to expectations, there was a decrease in accuracy with increasing levels of support. There was a significant decrease in accuracy from without compared to with support of DNN classifications (72.7% vs. 66.1%) and there was a significant tendency of decrease from with support of DNN classifications compared to with support of RISE saliency map explanations (66.1% vs. 59.1%) (Figure 4). For comparison, the accuracy of the DNN classifier within the experiment was 72.2%.

Figure 4

Mean and standard error of accuracy in target classification without support of DNN classifications, with support of DNN classifications, and support of RISE saliency map explanations.



There was a significant main effect of vehicle class *F*(5, 75) = 50.5, *p* < .001. There were significant differences in accuracy between T-72 (74.2%) and 2S3 (76.6%) and the other vehicle classes: BMP-2 (50.8%), BTR-80 (39.1%), ZSU-23 (55.2%), and Toyota Hilux (100%). The lower accuracy for ZSU-23 was due to confusion with all other vehicle classes, except Toyota Hilux. The lower accuracy for BMP-2 was mainly due to confusion with T-72, BTR-80, and 2S3. The lower accuracy for BTR-80 was mainly due to confusion with BMP-2 and 2S3.

There was a significant main effect of DNN classification correctness *F*(1, 5) = 55.8, *p* < .001. There was a significant decrease in accuracy from correct compared to incorrect DNN classification (72.7% vs. 59.2%).

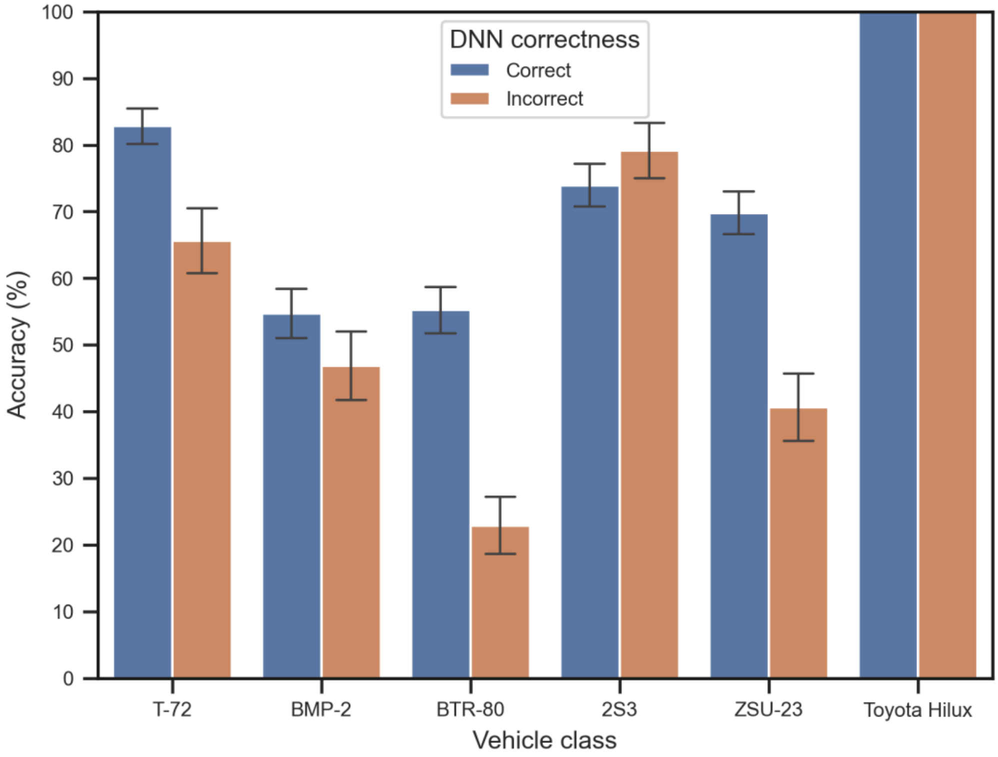
There was a significant main effect of vehicle resolution *F*(1, 15) = 15.5, *p* = .001. There was a significant increase in accuracy from low compared to medium resolution of vehicles (62.8% vs. 69.2%).

There were significant two-way interaction effects of vehicle class and each of the three other factors. There was a significant two-way interaction effect between vehicle class and support *F*(10, 15) = 2.95, *p* = .002. There was a significant decrease in accuracy from without compared to with support of RISE saliency map explanations for BTR-80 (54.7% vs. 35.2%) and ZSU-23 (69.5% vs. 51.6%). There was a significant tendency of decrease in accuracy from with support of DNN classifications compared to with support of RISE saliency map explanations for T-72 (80.5% vs. 53.3%).

There was a significant two-way interaction effect between vehicle class and DNN classification correctness *F*(5, 75) = 12.1, *p* < .001. There was a significant decrease in accuracy from correct compared to incorrect DNN classification for T-72 (82.8% vs. 65.6%), BTR-80 (55.2% vs. 22.9%), and ZSU-23 (69.8% vs. 40.6%) (Figure 5).

Figure 5

Mean and standard error of accuracy in target classification for the vehicle classes when the DNN classifications were correct and incorrect.



There was a significant two-way interaction effect between vehicle class and vehicle resolution *F*(5, 75) = 7.50, *p* < .001. There was a significant increase in accuracy from low compared to medium resolution of T-72 (67.2% vs. 81.2%) and BMP-2 (35.4% vs. 66.1%).

There was a significant tendency of a two-way interaction effect of support and correctness *F*(2, 30) = 2.59, *p* = .09. There was a significant tendency of decrease in accuracy from correct compared to incorrect DNN classification without support of DNN classifications (77.1% vs. 68.2%). There was significant decrease in accuracy from correct compared to incorrect DNN classification with support of DNN classifications (72.9% vs. 59.4%) and with support of RISE saliency map explanations (68.2% vs. 50.0%). Although some three- and four-way interaction effects were significant, they are not reported since they only provided limited information about the factors' effect on accuracy in target classification.

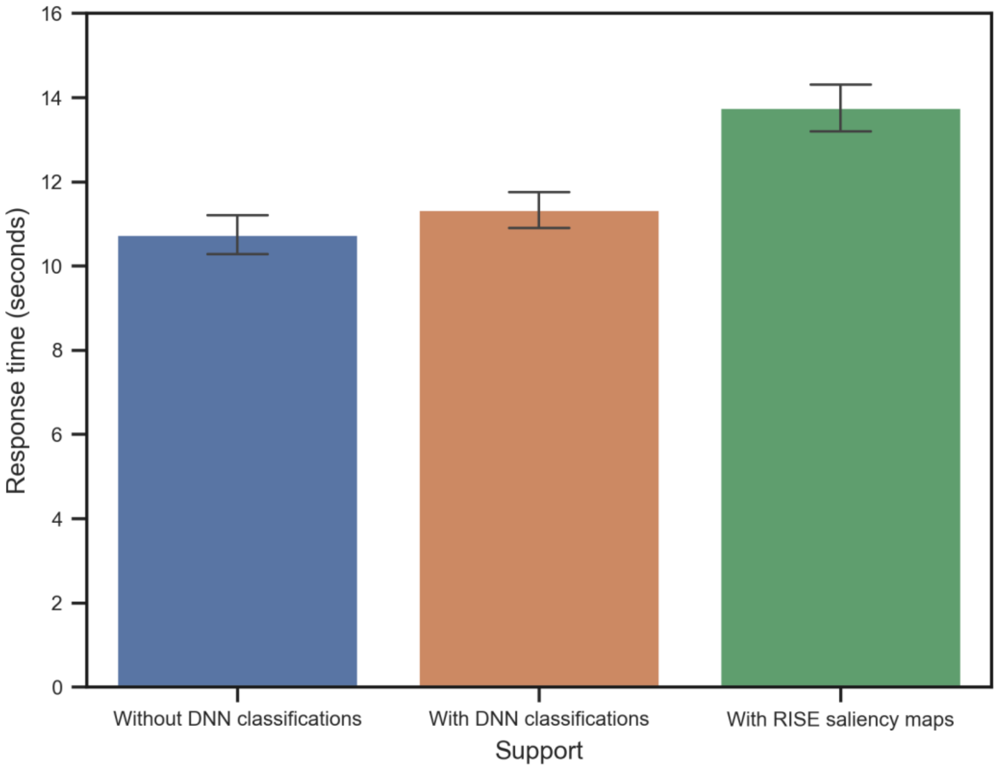
3.2 Response time for target classification

The response time for target classification was analyzed with a 3×6×2×2 repeated measures ANOVA of the factors support (without DNN classifications, with DNN classifications, and with RISE saliency map explanations), vehicle class (six classes), DNN classification correctness (correct vs. incorrect), and vehicle resolution (low vs. medium).

The analysis of response time for target classification showed significant main effects of all four factors. There was a significant main effect of support *F*(2, 30) = 4.45, *p* = .02. There was a significant tendency of increase in response time from without support of DNN classifications compared to support of RISE saliency map explanations (10.7 seconds vs. 13.8 seconds) and from with support of DNN classifications compared to with support of RISE saliency map explanations (11.3 seconds vs. 13.8 seconds) (Figure 6).

Figure 6

Mean and standard error of response time for target classification without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations.



There was a significant main effect of vehicle class *F*(5, 77) = 43.8, *p* < .001. There was significant increase in response time from Toyota Hilux (3.0 seconds) compared to the other vehicle classes: T-72 (11.5 seconds), BMP-2 (15.1 seconds), BTR-80 (14.0 seconds), 2S3 (12.8 seconds), and ZSU-23 (15.1 seconds). There was a significant increase in response time from T-72 compared to BMP-2 (11.5 seconds vs. 15.1 seconds) and there was a significant tendency of increase in response time from T-72 compared to ZSU-23 (11.5 seconds vs. 15.1 seconds).

There was a significant main effect of DNN correctness *F*(1, 15) = 7.15, *p* = .02. There was a significant increase in response time from correct compared to incorrect DNN classification (11.4 seconds vs. 12.5 seconds).

There was a significant main effect of vehicle resolution *F*(1, 15) = 26.6, *p* < .001. There was a significant decrease in response time from low compared to medium resolution of vehicles (12.9 seconds vs. 11.0 seconds).

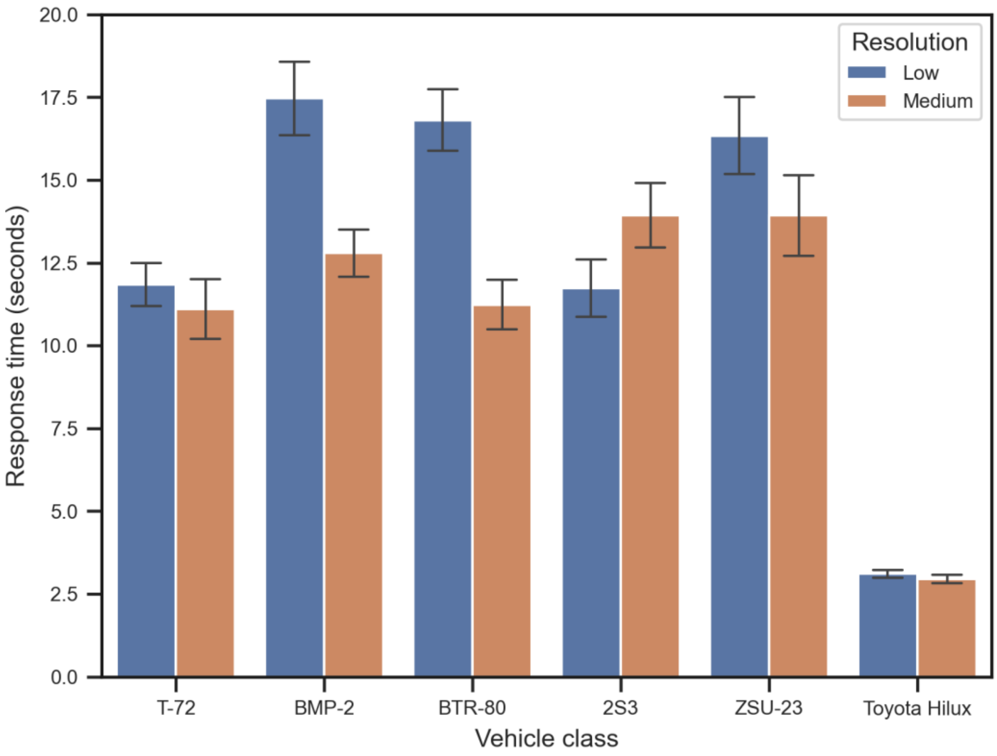
There were significant two-way interaction effects of vehicle class and each of the three other factors. There was a significant two-way interaction effect of vehicle class and support *F*(10, 150) = 3.36, *p* < .001. There was a significant increase in response time for T-72 from with support of DNN classifications compared to with support of RISE saliency map explanations (9.1 seconds vs. 14.0 seconds).

There was a significant two-way interaction effect between vehicle class and DNN correctness *F*(50, 75) = 6.37, *p* < .001. There was a significant increase in response time for ZSU-23 from correct compared to incorrect DNN classification (12.5 seconds vs. 17.8 seconds).

There was a significant two-way interaction effect between vehicle class and vehicle resolution *F*(5, 75) = 10.9, *p* < .001. There was a significant decrease in response time from low compared to medium resolution of BMP-2 (17.5 seconds vs. 12.8 seconds) and BTR-80 (16.8 seconds vs. 11.2 seconds) (Figure 7). There was a significant tendency of increase in response time from low compared to medium resolution of 2S3 (11.7 seconds vs. 13.9 seconds). Although some three- and four-way interaction effects were significant, they are not reported since they only provided limited information about the factors' effect on response time.

Figure 7

Mean and standard error of response time for target classification for the vehicle classes when the resolution of vehicles was low and medium.



3.3 Reliance on DNN classification

Reliance on the DNN classification was analyzed with a 3×4×2 repeated measures ANOVA of the factors support (without DNN classifications, with DNN classifications, and with RISE saliency map explanations), DNN classification (four vehicle classes), and DNN classification correctness (correct vs. incorrect). The factor DNN classification only included the vehicle classes 2S3, BMP-2, BTR-80, and ZSU-23 since these vehicle classes showed incorrect DNN classifications for all three levels of support. Reliance on the DNN classification is a misnomer for the condition without support of DNN classifications since classifications were not shown to the participants in this condition. However, the percentage of identical target classifications still served as a baseline for comparison with the other two conditions where DNN classifications were shown to the participants.

The analysis of reliance on the DNN classification showed significant main effects of all three factors. There was a significant main effect of support $F(2, 30) = 4.35$, $p = .02$. There was a significant increase in reliance on the DNN classification when it was shown compared to the baseline condition without support of DNN classifications. There was a significant increase in reliance from without compared to with support of DNN classifications (40.0% vs. 49.3%).

There was a significant main effect of DNN classifications $F(3, 35) = 8.05$, $p < .001$. There was a significant increase in reliance from BTR-80 (36.5%) compared to BMP-2 (54.9%) and 2S3 (48.1%).

There was a significant main effect of DNN classification correctness $F(1, 15) = 40.0$, $p < .001$. The reliance decreased significantly for correct compared to incorrect DNN classifications (63.4% vs. 28.0%).

There was a significant two-way interaction effect of support and DNN classification correctness $F(2, 30) = 12.8$, $p < .001$ (Figure 8). When the DNN classification was correct, there was a significant decrease in reliance from without support of DNN classifications compared to with support of RISE saliency map explanations (69.1% vs. 57.4%). When the DNN classification was incorrect, there was a significant increase in reliance from without compared to with support of DNN classifications (10.9% vs. 35.0%) and from without support of DNN classifications compared to with support of RISE saliency map explanations (10.9% vs. 37.9%). Although some additional two- and three-way interaction effects were significant, they are not reported since they only provided limited information about the factors’ effect on reliance.

3.4 Questionnaires

Participants’ subjective ratings of the questionnaires for mental workload, trust, and satisfaction were analyzed with separate repeated measures ANOVAs with factorial design.

3.4.1 Mental workload

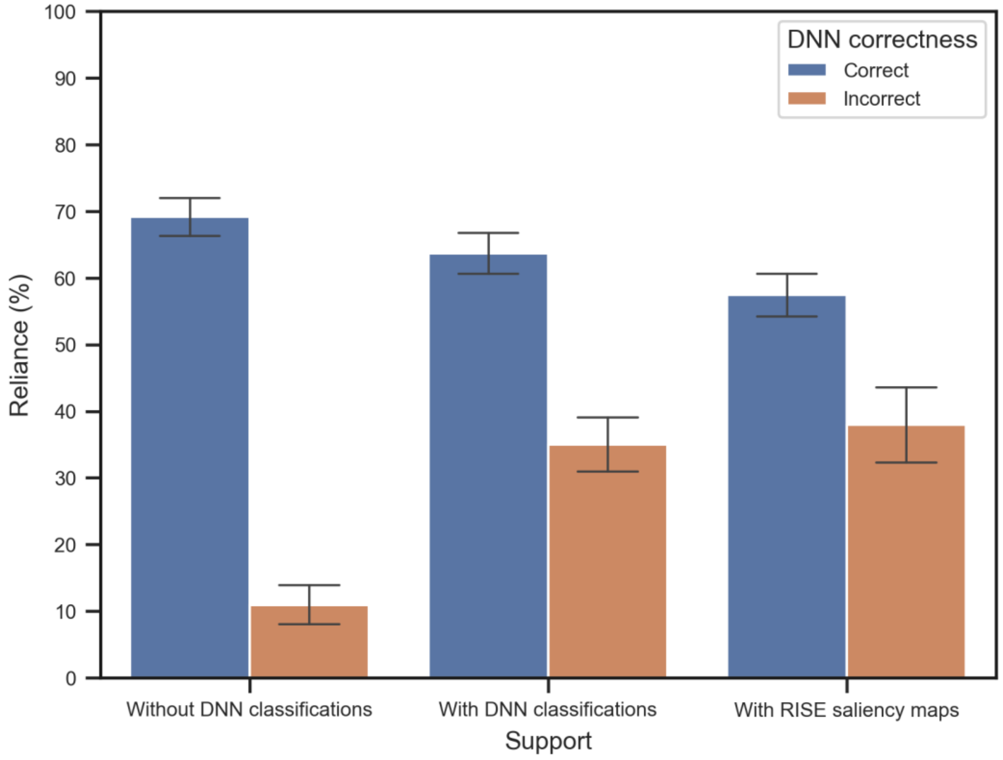
The subjective ratings of mental workload were analyzed with a one-way ANOVA of the factor support (without DNN classification, with DNN classification, and with RISE saliency map explanations). Ratings of the item Performance were reversed and the mean rating of workload items was calculated before analysis. The analysis did not show a significant effect of support.

3.4.2 Trust in DNN classifications and RISE saliency map explanations

The subjective ratings of trust in DNN classifications and RISE saliency map explanations were analyzed with a 3×6 repeated measures ANOVA of the factors support (with DNN classifications without RISE saliency map explanations, DNN classifications with RISE

Figure 8

Mean and standard error of reliance on the DNN classification when the classification was correct and incorrect for without support of DNN classifications, with support of DNN classifications, and with support of RISE saliency map explanations.



saliency map explanations, and with RISE saliency map explanations by themselves) and trust item (six items). Ratings of the trust item ‘I am wary of the automatic system’ were reversed before analysis.

The analysis of trust in DNN classifications and RISE saliency map explanations showed significant main effects of both factors. There was a significant main effect of support $F(2, 30) = 5.73$, $p = 0.008$. There was a significant decrease in trust from with support of DNN classifications with RISE saliency map explanations compared to with support of RISE saliency map explanations by themselves (3.2 vs. 2.4) (Figure 9). There was a significant tendency of decrease in trust from with support of DNN classifications without RISE saliency map explanations compared to with support of RISE saliency map explanations by themselves (3.1 vs. 2.4).

There was a significant main effect of trust item $F(5, 75) = 10.3$, $p < .001$. The mean ratings of the trust items was highest for ‘I like using the automatic system for decision making’ and lowest for ‘The automatic system is very reliable. I can count on it to be correct all the time.’ (Table 1). There was a significant decrease in subjective ratings from ‘I like using the automatic system for decision making’ and ‘I am confident in the automatic system. I feel that it works well’ compared to ‘I am wary of the automatic system’ and ‘The automatic system is very reliable. I can count on it to be correct all the time’. There was a significant decrease in subjective ratings from ‘The outputs of the automatic system are predictable’ compared to ‘The automatic system is very reliable. I can count on it to be correct all the time’. There was a significant tendency of decrease in subjective ratings from ‘I like using the automatic system for decision making’ compared to ‘I feel safe that when I rely on the automatic system, I will get the right answers’.

Figure 9

Mean and standard error of subjective ratings of trust in the DNN classifications without RISE saliency map explanations, DNN classifications with RISE saliency map explanations, and RISE saliency map explanations by themselves.

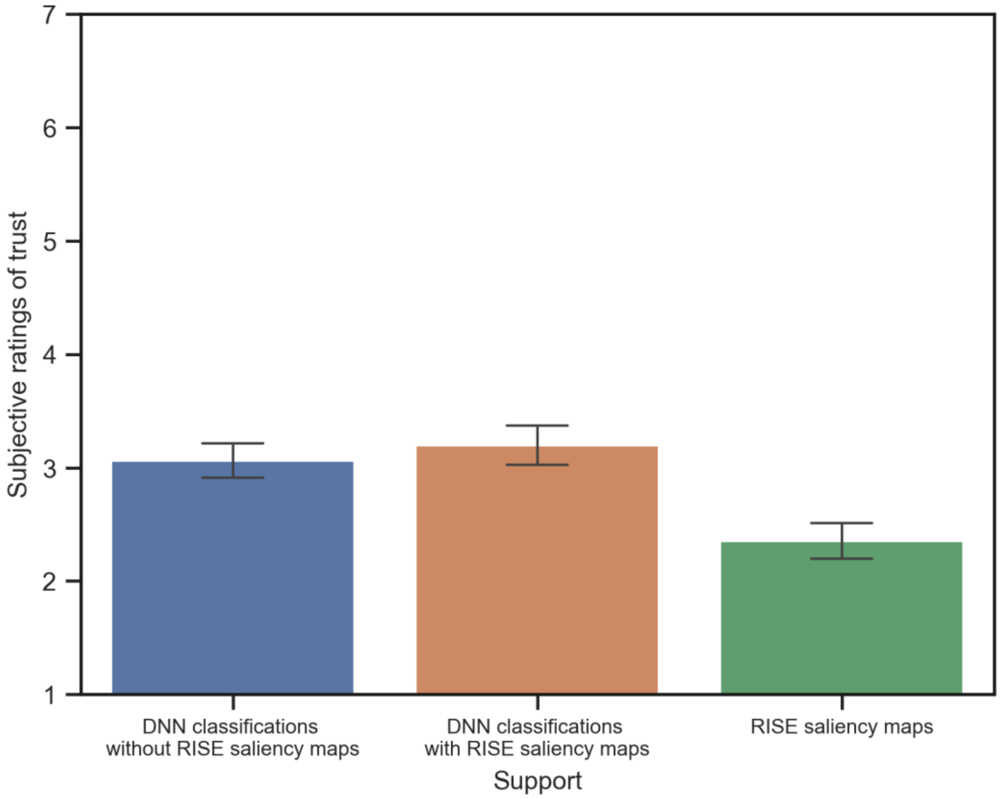


Table 1

Mean and standard error of subjective ratings of the trust items.

Trust item	Mean rating	Standard error
I like using the automatic system for decision making.	3.9	0.3
I am confident in the automatic system. I feel that it works well.	3.4	0.2
The outputs of the automatic system are predictable.	3.1	0.2
I feel safe that when I rely on the automatic system, I will get the right answers.	2.8	0.2
I am wary of the automatic system. (reversed)	2.1	0.2
The automatic system is very reliable. I can count on it to be correct all the time.	2.1	0.2

There was a significant two-way interaction effect of support and trust item *F*(10, 150) = 3.85, *p* < .001. There was a significant decrease in subjective ratings from with support of DNN classifications with RISE saliency map explanations compared to RISE saliency map explanations by themselves for ‘I like using the automatic system for decision making’ (4.6 vs. 2.6). There was a significant tendency of decrease in subjective ratings from with support of DNN classifications with RISE saliency map explanations compared to RISE saliency map explanations by themselves for ‘I feel safe that when I rely on the automatic system, I will get the right answers’ (3.2 vs. 2.2).

3.4.3 Satisfaction of RISE saliency map explanations

The subjective ratings of satisfaction of RISE saliency map explanations were analyzed with a one-way repeated measures ANOVA of the factor satisfaction item (six items).

The analysis did not show a significant effect of satisfaction item. The mean ratings of the satisfaction items was highest for ‘From the explanations, I understand how the automatic system works’ and lowest for ‘The explanations of the automatic system shows me how accurate it is’ (Table 2). There were no significant differences between the satisfaction items.

Table 2

Mean and standard error of subjective ratings of the satisfaction items.

Satisfaction item	**Mean rating**	**Standard error**
From the explanations, I understand how the automatic system works.	4.7	0.4
The explanations of how the automatic system works have sufficient detail.	4.4	0.5
The explanations of how the automatic system works are satisfying.	4.3	0.6
The explanations let me judge when I should trust and not trust the automatic system.	3.6	0.6
The explanations of the automatic system shows me how accurate it is.	3.6	0.5

4 Discussion

The results of the experiment only support one of the four hypotheses. Firstly, the results of the experiment support the hypothesis that accuracy in target classification is higher for images with higher vehicle resolution. However, the accuracy was only marginally higher. This was likely due to the small difference between the vehicle resolution intervals. Secondly, the results of the experiment did not support the hypothesis that support of DNN classifications improves accuracy in target classification. The accuracy in target classification was lower with support of DNN classifications compared to without support of DNN classifications. Thirdly, the results of the experiment did not support the hypothesis that support of RISE saliency map explanations improve accuracy in target classification. The accuracy in target classification was lower with support of RISE saliency map explanations of DNN classifications compared to with support of only DNN classifications. The main reasons for the lower accuracy with RISE saliency map explanations were disproportionate effects of the T-72 vehicle class and incorrect DNN classifications. The participants also trusted the DNN classifications more than the RISE saliency map explanations, particularly in terms of using the DNN classifications and their predictability. However, the participants were still moderately satisfied with the RISE saliency map explanations. Finally, the results of the experiment did not support the hypothesis that accuracy in target classification is higher with RISE saliency map explanations when the DNN classifications are incorrect. The differences in accuracy in target classification between incorrect and correct DNN classifications were similar both with and without support of RISE saliency map explanations. Kim et al. (2022) report similar results that saliency map explanations may not be distinct enough for users to detect incorrect predictions.

The experiment shows that the participants tried to use the DNN classifications and RISE saliency map explanations to improve accuracy in target classification. The participants' response time for target classification was longer with support of RISE saliency map explanations compared to without or with support of DNN classifications. Additionally, the participants relied more on the DNN classifications compared to the baseline condition without support of DNN classifications where participants could only use their own judgement. Although DNN classifications were not shown to the participants in the baseline condition, the percentage of identical target classifications still enabled comparison with the other two conditions where DNN classifications were shown to the participants. The increase in reliance with support of DNN classifications was likely due to the participants' difficulties in judging the reliability of the DNN classification, both with and without support of RISE saliency map explanations. Ideally, participants should rely fully on the DNN classification when it is correct and not at all when it is incorrect. Instead, the participants' reliance on correct DNN classifications decreased and their reliance on incorrect DNN classifications increased compared to the baseline condition without support of DNN classifications. The DNN classifications therefore undermined the participants' own judgements of target classification. Under-reliance on the DNN classification when it is correct and over-reliance when it is incorrect, decrease the accuracy in target classification with and without support of RISE saliency map explanations. Neither information from DNN confidence values, nor RISE saliency map explanations, were sufficient to improve the accuracy in target classification. The participants' difficulty of judging the reliability of the DNN classification explains their low trust in the DNN classifications both with and without support of RISE saliency map explanations. Participants' strong tendency to rely on the DNN classifications despite their low trust is in accordance with users' tendency to rely on automated decision aids (Dzindolet et al., 2003).

The experiment shows that the vehicle classes BMP-2, BTR-80, and ZSU-23, were more difficult to classify than T-72, 2S3, and Toyota Hilux. Only about half of the images of BMP-2, BTR-80, and ZSU-23 were classified correctly. These vehicle classes also required longer response time for classification, particularly for low resolution of the vehicles. BMP-2 and BTR-80 were mainly misclassified as vehicles with similar overall features, while difficult images for ZSU-23 made it look similar to all other vehicle classes except Toyota Hilux. The discriminating features for these vehicles were often insufficient for correct target classification, which was amplified when the DNN classification was incorrect and for low resolution of vehicles.

5 Conclusions

Target classification in UAV imagery is a challenging task where DNNs can improve accuracy in target classification. However, DNNs lack transparency for which image features are most important for the DNN's classification. This lack of transparency is a challenge for military applications where operators are ultimately responsible for all decisions due to the high risks of weapon engagements. Operators therefore also need explanations of DNN classifications to assess their reliability.

The conclusion of the experiment is that it is not trivial to create DNN classifications and explanations of DNN classifications that actually support operators' target classification. Contrary to expectations, support of DNN classifications and confidence values decrease the participants' accuracy in target classification compared to without support of DNN classifications. Further, the support of RISE saliency map explanations of the DNN classifications results in an additional decrease of the participants' accuracy in target classification. A likely reason for the participants' lower accuracy in target classification with these two types of support is the difficulty in assessing the DNN classifications' reliability. This results in under-reliance on the DNN classifications when they are correct and over-reliance when they are incorrect. The two types of support therefore undermine the participants' own judgment of target classifications. Other studies report similar negative effects of XAI that make participants more likely to follow the DNN classification or prediction (e.g. Zhang et al., 2020; Bansal et al., 2021; Nguyen et al., 2021).

6 Future work

Since RISE saliency map explanations were insufficient for the participants to assess the accuracy of DNN classifications, additional experiments should evaluate other promising XAI-approaches. Some examples of promising XAI-approaches are:

- Methods that use the internal parameters of DNNs to generate more precise saliency map explanations of what image features are most important for the DNN’s classification, such as Layer-Wise Relevance Propagation (LRP) (Montavon et al., 2019).
- Counterfactual explanations which highlight features that would change the DNN’s classification (e.g. Chou et al., 2022; Delany et al., 2023). Such explanations mimics human explanations of causality for choosing one option over another.
- Example-based explanations that generate examples from the training set that are most similar to the input image in the DNN’s higher level representation of images (Jeyakumar et al., 2020). Results show that users prefer such examples over saliency map explanations.
- Visual correspondence-based explanations that are similar to example-based explanations, but compare image patches instead of the whole image (Nguyen et al., 2021). Results show that such explanations help users reject incorrect DNN classifications, which results in higher performance than either users or DNN alone.
- Concept-based explanations that explain DNN classifications using human-understandable attributes or abstractions that resemble human explanations (Poeta et al., 2023).
- Natural language explanations that mimics human explanations (e.g. Cambria et al., 2023).

Additionally, any XAI-approach for target classification may benefit from DNNs that are aligned with human visual strategies for object classification (Fel et al., 2022). Additional experiments should also consider how XAI-approaches may be combined with alternative methods for measuring DNN uncertainty to inform participants about the reliability of DNN classifications (e.g. Astorga et al., 2023; Malmström et al., 2024). With the considerable research in XAI it is important to continue investigations and evaluations of whether such options actually support participants and increase their accuracy in military target classifications.

References

Adebayo, J., Muelly, M., Abelson, H., & Kim, B. (2022). Post hoc explanations may be ineffective for detecting unknown spurious correlation. *In Proceedings of the International Conference on Learning Representations (ICLR) 2022*.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 275-285).

Astorga, A., Hsieh, C., Madhusudan, P., & Mitra, S. (2023). Perception Contracts for Safety of ML-Enabled Systems. In *Proceedings of the ACM on Programming Languages*, *7*(OOPSLA2), 2196-2223.

Bachmann, T., & Francis, G. (2014). Visual masking: Studying perception, attention, and consciousness. In T. Bachmann & G. Francis (Eds,), *Visual Masking* (pp. 1–108). Elsevier.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, *60*(1), 103111.

Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59-83.

Colin, J., Fel, T., Cadène, R., & Serre, T. (2022). What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, *35*, 2832-2845.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371*.

Delaney, E., Pakrashi, A., Greene, D., & Keane, M. T. (2023). Counterfactual explanations for misclassified images: How human and machine explanations differ. *Artificial Intelligence*, *324*, 103995.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697-718.

Ehsan, U., & Riedl, M. O. (2021). Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480*.

Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, *35*, 9432-9446.

Field, A. (2024). *Discovering statistics using IBM SPSS Statistics* (sixth ed.). Sage.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745-e750.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2019). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can I explain this to you? An empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems, 33,* 4211–4222.

Kim, S. S., Meister, N., Ramaswamy, V. V., Fong, R., & Russakovsky, O. (2022). HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision* (pp. 280-298). Springer Nature Switzerland.

Kullab, S. (2023). *Ukraine is building an advanced army of drones. For now, pilots improvise with duct tape and bombs.* https://apnews.com/article/drones-ukraine-war-russia-innovation-technology-589f1fc0e0db007ea6d344b197207212.

Li, X. H., Shi, Y., Li, H., Bai, W., Cao, C. C., & Chen, L. (2021). An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3200-3208).

Lif, P., Näsström, F., Bissmarck, F., & Allvar, J. (2018). User performance for vehicle recognition with visual and infrared sensors from an unmanned aerial vehicle. In *Proceedings of International Conference on Human-Computer Interaction* (pp. 295-306). Springer.

Lif, P., Näsström, F., Karlholm, J., Allvar, J. (2021). *Värdering av AI-algoritm för luftburen termisk sensor* [*Evaluation of AI-algorithm for Airborne Thermal Sensor]* (In Swedish) (FOI-R--5200--SE). Swedish Defence Research Agency.

Luotsinen, L. J., Oskarsson, D., Svenmarck, P. & Wickenberg Bolin, U. (2019). *Explainable Artificial Intelligence: Exploring XAI techniques in Military Deep Learning Applications* (FOI-R--4849--SE). Swedish Defence Research Agency.

Malmström, M., Skog, I., Axehill, D., & Gustafsson, F. (2024). Uncertainty quantification in neural network classifiers—A local linear approach. *Automatica, 163,* 111563.

Mittal, P., Sharma, A., & Singh, R. (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image and Vision Computing*, 104046.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K. R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 193-209). Springer Nature.

Nguyen, G., Kim, D., & Nguyen, A. (2021). The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, *34*, 26422-26436.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018 (BMVC 2018)* (p. 151), Northumbria University, Newcastle, UK, September 3-6, 2018.

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., & Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11443-11452).

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., & Baralis, E. (2023). Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*.

Rubio, L. (2020). *Infantry soldiers test short-range reconnaissance unmanned aircraft amid COVID-19.* Retrieved from https://www.army.mil/article/238277/infantry_ soldiers_test_short_ range_reconnaissance_unmanned_aircraft_amid_covid_19.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Stanchi, O., Ronchetti, F., Quiroga, F. (2023). The implementation of the RISE algorithm for the Captum framework. In Naiouf, M., Rucci, E., Chichizola, F., & De Giusti, L. (Eds.), *Cloud Computing, Big Data & Emerging Topics*. JCC-BD&ET 2023. Communications in Computer and Information Science, vol 1828. Springer, Cham. https://doi.org/10.1007/978-3-031-40942-4_7.

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018). Possibilities and challenges for artificial intelligence in military applications. In *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting* (pp. 1-16). NATO-STO.

Tsunakawa, H., Kameya, Y., Lee, H., Shinya, Y., & Mitsumoto, N. (2019). Contrastive relevance propagation for interpreting predictions by a single-shot object detector. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *arXiv preprint arXiv:2006.00093*.

Wu, X., Li, W., Hong, D., Tao, R., & Du, Q. (2021). Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, *10*(1), 91-124.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295-305).

Appendix. Questionnaires

Subjective Workload

These questions concern your subjective workload during the vehicle classification task. Please indicate your preferred answer regarding each statement on a scale from 1 to 10.

Mental Demand – How mentally demanding was the task?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Physical Demand – How physically demanding was the task?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Temporal Demand – How hurried or rushed was the pace of the task?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Performance – How successful were you in accomplishing what you were asked to do?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Effort – How hard did you have to work to accomplish your level of performance?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Frustration – How insecure, discouraged, irritated, stressed, and annoyed were you?

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Very Low									Very High

Trust in Classifier

The following statements concern your trust in the Classifier. Please indicate your preferred answer regarding each statement on a scale from 1 to 7.

1. I am confident in the Classifier. I feel that it works well.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

2. The outputs of the Classifier are predictable.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

3. The Classifier is very reliable. I can count on it to be correct all the time.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

4. I feel safe that when I rely on the Classifier, I will get the right answers.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

5. I am wary of the Classifier.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

6. I like using the Classifier for decision making.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

Trust in Heatmap explanations

The following statements concern your trust in the Heatmap explanations for the Classifier’s classifications. Please indicate your preferred answer regarding each statement on a scale from 1 to 7.

1. I am confident in the Heatmap explanations. I feel that they work well.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

2. The Heatmap explanations are predictable.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

3. The Heatmap explanations are very reliable. I can count on them to be correct all the time.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

4. I feel safe that when I rely on the Heatmap explanations, I will get the right answers.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

5. I am wary of the Heatmap explanations.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

6. I like using the Heatmap explanations for decision making.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

Satisfaction of Heatmap explanations

The following statements concern your satisfaction with the Heatmap explanations for the Classifier's classifications. Please indicate your preferred answer regarding each statement on a scale from 1 to 7.

1. From the explanations, I understand how the Classifier works.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

2. The explanations of how the Classifier works are satisfying.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

3. The explanations of how the Classifier works have sufficient detail.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

4. The explanations of the Classifier shows me how accurate it is.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

5. The explanations let me judge when I should trust and not trust the Classifier.

○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Strongly disagree						Strongly agree

FOI
Defence Research Agency
SE-164 90 Stockholm

Phone: +46 8 555 030 00
Fax: +46 8 555 031 00

www.foi.se