Transparent Assessment of Automated Human Detection in Aerial Images via Explainable AI

Sara Narteni *, Maurizio Mongelli *
Italian National Research Council - Institute of Electronics, Information Engineering and Telecommunications (CNR-IEIIT), Genoa, 16152, Italy

Joachim Rüter†, Christoph Torens†, Umut Durak †
German Aerospace Center (DLR) - Institute of Flight Systems, Brunswick, 38108, Germany

The widespread adoption of Artificial Intelligence (AI)-based software technologies supporting Unmanned Aircraft Systems (UASs) demands new validation methods to ensure that these systems operate safely and reliably. This paper investigates the role of eXplainable AI (XAI) and, in particular, of rule-based models in monitoring the performance of a deep learning-based human detector from aerial images. Starting from several image attributes extracted from the images and information about the performance of the detection model, decision rules are extracted to map the image attributes onto the performance quality. Besides shedding light on the logic of the humans detection successes and failures, these rules can serve as a performance monitor at runtime, by triggering alerts in case input images do not satisfy them. The obtained rules have been adopted to filter out inputs associated with bad performance, showing improved precision and recall with respect to the original model, thus opening the road to promising future developments.

I. Introduction

Thanks to the fast-paced rising of sensor technologies and Artificial Intelligence (AI), Unmanned Aircraft Systems (UAS) are finding application in several fields, including urban traffic management[1], environmental monitoring [2], video surveillance [3], smart agriculture [4] and many others [5]. One of the most investigated tasks is object detection, which leverages the boosting of advanced AI algorithms and infrastructures, and in particular of Deep Neural Networks (DNN), to come up with models characterized by very high-performance capabilities and relatively low computational costs [6].

While being a key enabler for UAS, AI also brings up new fundamental challenges, residing in its *verification and validation*, ensuring that the autonomous decisions made by the AI models do not cause harm to humans or damage to the surrounding environment. The AI safety assurance problem is indeed part of a wider, multi-faceted, and multi-disciplinary paradigm, being referred to as *Trustworthy AI* (TAI), and recently governed by institutions such as the European Commission Ethics Guidelines [7] or most recent regulations (see, e.g., the EU AI Act [8]). Focusing on avionics, field regulations also arise in the community such as the European Union Aviation Safety Agency (EASA) [9–11] and others [12].

Such certification processes pose many challenges, especially when humans are involved, e.g., in emergency medicine scenarios [13], search and rescue [14] or dropping goods, since failures of AI-guided detection systems might result in severe harms to people. Despite reaching promising results, DNN-based human detection models have a black-box nature, preventing the possibility of understanding why the model generated its outcomes and, subsequently, analyzing the reasons for correct results and failures. In this context, and in compliance with TAI principle of *transparency*, the branch of *eXplainable AI (XAI)* comes to help, offering a set of techniques to either design intrinsically interpretable models or to provide explanations to black-boxes [15].

A. Contribution

In an attempt to address these issues, this paper investigates the innovative use of rule-based classifiers as a transparent validation tool of a deep learning (DL) model for human detection in aerial images. More specifically, the

*Researcher, CNR-IEIIT, {sara.narteni, maurizio.mongelli}@cnr.it
†Researcher, Institute of Flight Systems, {joachim.rueter, christoph.torens, umut.durak}@dlr.de

objective is to obtain a set of interpretable *if-then* rules characterizing the space of image features associated with a good or bad performance of that model.

Besides shedding light on the logic of the human detection successes and failures, these rules can then serve as a performance monitor at runtime, by triggering alerts in case input images do not satisfy them. This can help identify corner cases of human detection, where the performance of the model is no longer guaranteed. For example, see Figure 1, where unusual light conditions like darkening hinder a correct model performance.

(a) Image from HERIDAL dataset with *correct* human detection

(b) Darkened and noisy version of the same image: the person on the left is *not detected*

Fig. 1 Examples of an image from HERIDAL dataset [16], where the AI performs well (left) or fails (right). Green boxes symbolize ground truth, magenta boxes symbolize predictions

The overall idea of our approach is shown in Figure 2. The green arrow highlights the main concept of this paper, that is, combining well-established methodologies from object detection and XAI, and making them collaborative, in the sense that the latter can serve as a monitoring and improvement tool for the first. When dealing with high-dimensional data like images, however, it is not trivial to individuate a representative set of features with good discriminant ability between classes. Therefore, we started working on the feature extraction and selection phase and individuated a set of variables useful to train a rule-based classifier that has satisfying performance. This is our starting point, and the next investigations will be devoted to verifying the viability of the proposed approach. In the following, we describe the use case of reference, the main methodological aspects of rule-based classification, some preliminary results, and the challenges we will be trying to address through further experimentation.

II. Related Work

Over the last years, using DL-based methods to improve the perception capabilities of UASs received more and more attention [3, 17, 18]. A problem arises as the current aviation certification processes cannot be directly applied to the data-driven learning processes of DL. Regulatory bodies like the EASA and the Federal Aviation Authority (FAA) have recognized this problem and are actively investigating new certification avenues [10, 11, 19, 20].

As a result, some of the new objectives in the EASA guidelines require some form of monitoring for the inputs of the machine learning component. Originally coming from the automotive domain [21–23], in this context EASA

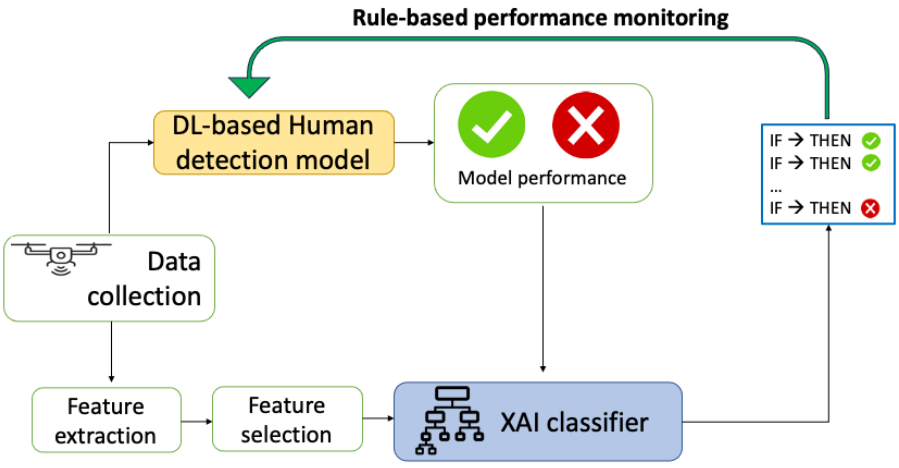

Fig. 2 Flowchart of the proposed idea. A deep learning-based human detection model is applied to images recorded from a UAS. Image properties extracted from the images, combined with information on good or bad detection performance, are fed to a XAI classifier, generating a set of rules for monitoring

introduced the concept of operational design the main (ODD). The idea of the ODD is to define "Operating conditions under which a given AI/ML constituent is specifically designed to function as intended" [11]. Such operating conditions can be: time of day of the operation, weather, illumination, or brightness of an image [24]. The idea of the ODD is to ensure that all input images are inside of the allowed range of the operating conditions in that the AI constituent is specified to work correctly. However, the characterization of the ODD can be quite complex [25]. The upcoming process standards EUROCAE ED 324 / SAE ARP 6983 will give more guidance on the development and certification of AI. A partially compliant concept is shown in [26], e.g. detailing ODD definition and and data design.

Furthermore, there is the concept of Out-of-Distribution (OOD) for input images of machine learning models. The idea is again to ensure inputs, but looking at the distributions of specific parameters for the training data [27, 28]. For example, such a parameter could be the brightness of the image or the altitude of the UAS [29]. Then, the distribution of the brightness of training images is analyzed and new input is compared to this distribution. Still, there are arguments that, even looking at the distributions of parameters, is not sufficient for input monitoring [30].

Another approach to improve the trustworthiness of DL systems is using XAI techniques [31, 32]. XAI literature is commonly categorized into two broad ways of performing explainability: on the one hand, *post-hoc* techniques [33] provide some form of interpretation (e.g., via rules, feature importance plots, saliency maps, etc.) to black-box predictions; on the other hand, *interpretable-by-design* techniques aim at training fully transparent models [15, 34].

III. Use case definition

A. Dataset

In this work, we consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, b_i)|i = 1, \ldots, N\}$ composed of $N = 1924$ annotated images $\boldsymbol{x}_i$ each containing M_i bounding boxes $b_i = \{(x_{ij}, y_{ij}, w_{ij}, h_{ij})|j = 1, \ldots, M_i\}$ of humans. As a basis, the publicly available *PeopleOnGrass* dataset [35] is used. It contains images of humans on mostly grassy areas taken from various angles and altitudes. We subsample the dataset to contain images of humans taken at altitudes between 4 and 70 m similar to [29]. Furthermore, we center-crop the images to be of size 2160x2160 pixels and resize them to 1080x1080 pixels to reduce the computational load of the object detection model. Samples of images from the dataset are shown in Figure 3.

Fig. 3 Examples of images taken from the PeopleOnGrass dataset [35]. Images are resized as described in Section III.A

B. Human detection model

We first build a human detection system based on a very well-established YOLOv7 [36] object detection model h, by using 750 images for training, 749 for validation, and the remaining 425 for testing. The model is trained for 300 epochs with a batch size of 16 using the standard training hyperparameters as recommended in the used implementation. For evaluation, the model with the best result on the validation dataset is used. This model performed sufficiently well, achieving the following values of mean Average Precision at an Intersection over Union threshold of 0.5 (mAP@0.5): 0.91 on the training dataset, 0.85 on the validation set, and 0.86 on the test data.

C. Classification problem definition

Given the trained human detection model, the next step is to define: i) binary labels y_i for each image in $\mathcal{D}$, expressing *whether* h performs correctly or not; ii) a set of features able to properly describe the images of the dataset, which is required for rule-based classification.

Concerning the label definition, we decided to take the most cautious approach and set

$$y_i = \begin{cases} 1 & \text{if } \textit{all} \text{ people are well detected through } h, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, we consider the performance of the model correct when all M_i humans present in the image $\boldsymbol{x}_i$ are correctly recognized by the model h.

Feature extraction from the images involves computing several numerical indicators to put into evidence useful characteristics of the images. In our case, we extracted common image characteristics such as brightness, saturation, entropy, sharpness, and edges. Furthermore, we used the Python library *Athec* [37], which provides capabilities to calculate a wide range of color information through statistics (e.g., mean, median, standard deviations, quartiles, etc.) in several color spaces like RGB, HSV, HSL and L*a*b. Overall, we extract a set of $N_f = 208$ features in total, which compose feature vectors $\boldsymbol{z}_i \in \mathbb{R}^{N_f}$ associated to image $\boldsymbol{x}_i$. Therefore, a *binary classification dataset* is now defined as

$$\mathcal{D}_2 = \{(\boldsymbol{z}_i, y_i) | i = 1, \ldots, N\},$$

being suitable for studying the performance of model h through rule-based classification, whose fundamentals are given in the next Section.

IV. Rule-based classification

A. Notation

Rule-based classifiers belong to the XAI branch of *interpretability by design*, describing machine learning models that provide their decisions through sets of interpretable rules, i.e., rulesets $\mathcal{R} = \{r_k\}_{k=1}^{N_r}$. Each rule r_k is expressed

in the form [38]: **if** $premise$ **then** $consequence$. The *premise* part is a logical conjunct of conditions on the input features, i.e., formally:

$$premise(r_k) = \bigwedge_{i_k=1_k}^{N_k} c_{i_k}$$

Each rule has a set of N_k conditions c_{i_k}, each referring to a variable z_j and corresponds to an interval that can be bounded, only lower-bounded or only upper-bounded:

1. $z_j \geq l_{i_k}$
2. $l_{i_k} \leq z_j \leq u_{i_k}$
3. $z_j \leq u_{i_k}$

where l_{i_k} and u_{i_k} are proper *numerical thresholds* learned by the classifier. The *consequence* part expresses the target class $\hat{y}_k \in \{0, 1\}$ predicted by the rule.

The Logic Learning Machine (LLM) [39] is an example of a classification model of this kind, and the one we consider in this work. The next Section will thus provide the fundamentals of this method.

B. Logic Learning Machine

In short LLM, it is a rule-based classifier, designed as an evolution of Switching Neural Networks [40] by Rulex Innovation Labs*. The rule learning process follows three steps: 1) a *discretization* of the feature space and a mapping to a Boolean lattice; 2) the identification of groups of points (called implicants) in the Boolean space, associated to the output classes, through a technique called *shadow clustering* [41]; 3) a *rule generation* phase, where *if-then* rules are retrieved from the implicants clusters by converting them to the original space, and eventually combined into a set of intelligible rules. The LLM rule generation thus follows an *aggregate-and-conquer* approach, resulting in rules that can overlap, i.e., the same sample may cover multiple rules.

C. Rule evaluation

The predictive ability of each rule r_k of the model can be evaluated by two metrics, namely the covering $C(r_k)$ and error $E(r_k)$, commonly known as True Positive Rate and False Positive Rate of the rule, respectively. They are defined as follows:

$$C(r_k) = \frac{TP(r_k)}{TP(r_k) + FN(r_k)} \tag{2}$$

$$E(r_k) = \frac{FP(r_k)}{TN(r_k) + FP(r_k)} \tag{3}$$

where $TP(r_k)$ and $FP(r_k)$ are defined as the number of images that correctly or wrongly satisfy rule r_k, while $TN(r_k)$ and $FN(r_k)$ represent the number of samples correctly or wrongly not satisfy r_k, respectively. The combination of these metrics gives the *rule relevance*:

$$R(r_k) = C(r_k) \cdot (1 - E(r_k)). \tag{4}$$

Overall, covering and relevance, can thus be considered as good metrics to evaluate how well a rule can generalize to unseen data, by measuring the portion of points correctly covered by the rule.

D. Class label assignment

Once those rules are generated, they can be used to make inference on unseen points $\tilde{z}$, thus assigning a label $\hat{y}$ to them. For the LLM, this stage is performed as follows. Consider the set of rules $\mathcal{R}^y_{\tilde{z}}$, verified by $\tilde{z}$ and predicting label y, and let $\mathcal{R}^y$ be the set of all rules of the model predicting class y. Then, label $\hat{y}$ is assigned to $\tilde{z}$ by solving the following problem:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left(\frac{\sum_{r \in \mathcal{R}^y_{\tilde{z}}} R(r)}{\sum_{r \in \mathcal{R}^y} R(r)} \right). \tag{5}$$

Hence, rule relevance also has an important role in determining the inference results. Also, considering Eq. 5, the LLM model can be evaluated as any machine learning method, such as using a confusion matrix or other related metrics.

*https://www.rulex.ai/

V. Results

The LLM rule generation model is applied to dataset $\mathcal{D}_2$, by maintaining the same train/test/validation split used for the human detection phase. Also, a χ^2 independence test was carried out to determine the statistical significance of the rules. However, in analyzing and validating the first results, a problem linked to the high-dimensionality of the dataset

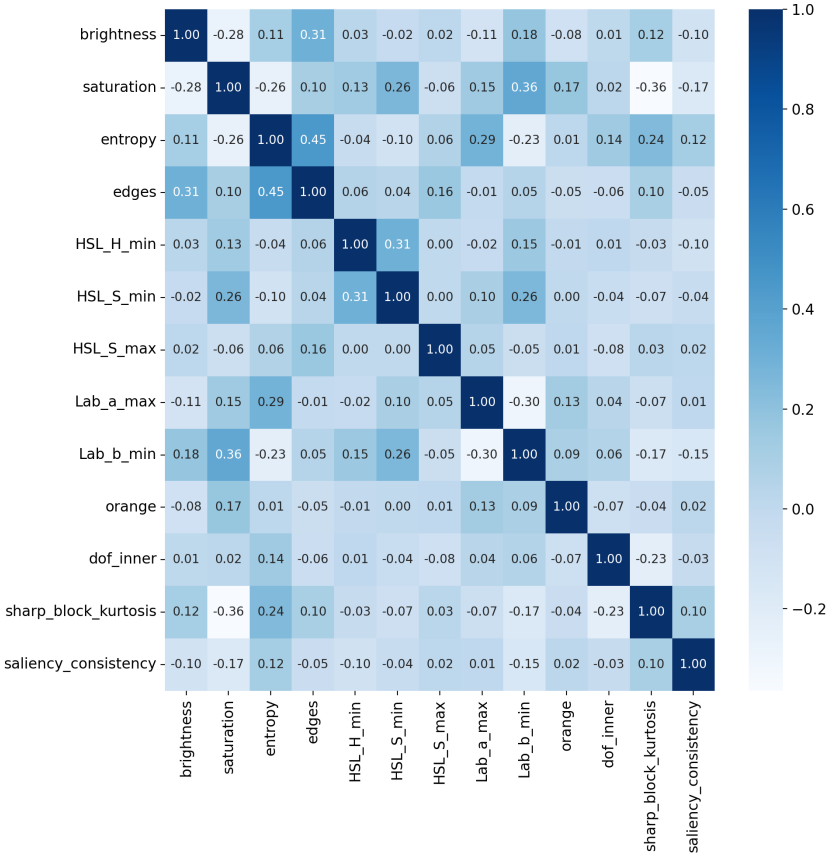

Fig. 4 Correlation matrix after feature selection via Pearson's correlation coefficient (threshold 0.5).

emerged. Indeed, while performing sufficiently well, with $TPR = 0.89$ and $FNR = 0.11$, on the target class $y = 1$ (i.e., correct human detection), we noticed that rules were often very long (i.e., with a high number of conditions) and, most importantly, not stable at random shuffles of the input samples. For this reason, we investigate the impact of a feature selection process.

A. Feature Selection

Pearson's correlation was computed among all 208 features, resulting in many pairs of features having absolute correlation values over 0.5. We thus performed a feature filtering by dropping all these features and keeping only the 13 variables with an absolute correlation coefficient < 0.5 between each other, as displayed in Figure 4. It can be observed that the correlation analysis preserves main features like *brightness, saturation, entropy, edges, saliency_consistency*, but also less intuitive ones such as *orange, Lab_a_max, Lab_b_min*.

B. Obtained Rules

After training the LLM model on the restricted set of features, we interestingly observed that, despite the drop of many variables, the overall performance did not significantly change, achieving a set of 12 rules (after statistical validation test) that scored $TPR = 0.86$, $TNR = 0.70$, $FNR = 0.14$, and $FPR = 0.30$ on the test data. And, in this case, these rules (at least those with larger covering) were approximately the same when randomly shuffling the rows of the training data.

Table 1 Performance comparison of the DL-based human detection model applied to all the original images versus the same model applied on the subset of images selected via rules.

		#images	#detections	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Training	Original	750	2651	0.96	0.90	0.91	0.55
	Rule-based filtering	527	1671	0.97	0.97	0.98	0.61
Validation	Original	749	2675	0.94	0.86	0.85	0.45
	Rule-based filtering	511	1591	0.97	0.96	0.96	0.54
Test	Original	425	1511	0.95	0.86	0.86	0.46
	Rule-based filtering	278	893	0.97	0.96	0.96	0.54

An example of two top-covering rules predicting correct human detections ($y = 1$) is given below:

1. **if** ($brightness \leq 0.551633 \wedge$
   $saturation > 0.195213 \wedge$
   $0.178954 < edges \leq 0.916329 \wedge$
   $Lab\_a\_max \leq 165 \wedge$
   $Lab\_b\_min > 79 \wedge$
   $dof\_inner \leq 1.668798 \wedge$
   $sharp\_block\_kurtosis > -0.996393$) **then** $y = 1, C = 0.46,\ E = 0.04$
2. **if** ($0.219320 < saturation \leq 0.537747 \wedge$
   $entropy \leq 0.942299 \wedge$
   $0.772010 < edges \leq 0.910104$) **then** $y = 1, C = 0.41,\ E = 0.05$

The covering values over 40% denote that, for the class $y = 1$, the LLM managed to individuate good descriptors of the class.

C. Human detection after rules application

All rules obtained have been used to filter the inputs of the DL model, by selecting only those images that satisfied the rules predicting the correct prediction class $y = 1$. This resulted in removing a portion of about 30-35% of the original images. The human detection model was then tested on this subset, and its performance in terms of precision, recall, mAP@0.5, and mAP@0.5:0.95 was calculated and compared to the performance on the original images. Table 1 shows the obtained results, for all training, validation, and test data.

We can observe how the filtering effectively improves the quality of the human detection, since larger values are registered for each of the considered metrics. Notably, recall increases by 7-10 percent points in all data portions, which means reducing the rate of people not being correctly detected, by avoiding to use the model on those images that most probably lead to missed detections. The metric mAP@0.5 also considerably improves, suggesting a better ability in recognizing people and correctly locate them, reducing false positives and negatives, which is also reflected in the good balance achieved between precision and recall. Finally, the detection task becomes more challenging when evaluating the mAP with IoU thresholds in the [0.5,0.95] interval: nevertheless, rule-based filtering still manages to significantly improve the original performance.

VI. Conclusions and Future Work

In this paper, we proposed the innovative application of a XAI model as a performance monitoring tool for a DL-based people detector fed with aerial images of humans. In our concept (see I.A), generated rules serve as an input montior for the DL model. The generated rules have been applied as a filter for model inputs, revealing that such a rule-guided image selection effectively improves the detection quality.

This is, however, just a starting point for a fully trustworthy-by-design solution. The effectiveness and generalizability of such an approach, in fact, requires a much deeper investigation, as rules themselves arise from a machine learning model that, even though interpretable, is subject to uncertainty that needs to be handled before being able to use the rules in practice. So, how to avoid that the errors of rule generation further propagate and reflect in the human detection

performance monitoring? Moreover, DNN-based object detection tasks often involve long training processes by using images from different sources, either real or even synthetic. And rule generation from real versus synthetic data is a matter of discussion around the kind of knowledge one can derive from rules, e.g., if rules on real versus synthetic data do not match, would the latter have to be considered 'wrong'? Or would it mean that new plausible factors are being discovered? Leveraging on the first performance evaluation carried out in this paper, future research will thus attempt to answer such questions.

Acknowledgments

This work was partially funded by Future Artificial Intelligence Research (FAIR) project, Italian Recovery and Resilience Plan (PNRR), Spoke 3 - Resilient AI. The work was also partially supported by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028.

References

[1] Srivastava, S., Narayan, S., and Mittal, S., "A survey of deep learning techniques for vehicle detection from UAV images," *Journal of Systems Architecture*, Vol. 117, 2021, p. 102152.

[2] Tang, G., Ni, J., Zhao, Y., Gu, Y., and Cao, W., "A Survey of Object Detection for UAVs Based on Deep Learning," *Remote Sensing*, Vol. 16, No. 1, 2023, p. 149.

[3] Mittal, P., Singh, R., and Sharma, A., "Deep learning-based object detection in low-altitude UAV datasets: A survey," *Image and Vision computing*, Vol. 104, 2020, p. 104046.

[4] Sassu, A., Motta, J., Deidda, A., Ghiani, L., Carlevaro, A., Garibotto, G., and Gambella, F., "Artichoke deep learning detection network for site-specific agrochemicals uas spraying," *Computers and Electronics in Agriculture*, Vol. 213, 2023, p. 108185.

[5] Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N. S., Khreishah, A., and Guizani, M., "Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges," *Ieee Access*, Vol. 7, 2019, pp. 48572–48634.

[6] Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X., "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, Vol. 30, No. 11, 2019, pp. 3212–3232.

[7] High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," Report, European Commission, Brussels, Apr. 2019. URL https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[8] Act, A. I., "EU AI Act, EUR-Lex - 52021PC0206," , 2024. URL https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-la.

[9] EASA AI Task Force, "Concepts of Design Assurance for Neural Networks CoDANN," Standard, European Union Aviation Safety Angency, Daedalean, AG, Mar. 2020. Also available as https://www.easa.europa.eu/sites/default/files/dfu/EASA-DDLN-Concepts-of-Design-Assurance-for-Neural-Networks-CoDANN.pdf.

[10] European Aviation Safety Agency (EASA), Deadalean AG, "Concepts of Design Assurance for Neural Networks (CoDANN) 2," Tech. rep., 2021. Also available as https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann-ii.

[11] EASA, "EASA Concept Paper: guidance for Level 1 & 2 machine learning applications Issue 02," , Mar. 2024. URL https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2.

[12] Schopferer, S., Donkels, A., Torens, C., Benders, S., Schirmer, S., Funke, A., and Dauer, J. C., "Machine Learning Applications in Unmanned Aviation: Operational Risks and Certification Considerations," *DEEL Workshop: Machine Learning in Certified Systems*, 2021.

[13] Carrillo-Larco, R. M., Moscoso-Porras, M., Taype-Rondan, A., Ruiz-Alejos, A., and Bernabe-Ortiz, A., "The use of unmanned aerial vehicles for health purposes: a systematic review of experimental studies," *Global health, epidemiology and genomics*, Vol. 3, 2018, p. e13.

[14] Golcarenarenji, G., Martinez-Alpiste, I., Wang, Q., and Alcaraz-Calero, J. M., "Efficient real-time human detection using unmanned aerial vehicles optical imagery," *International Journal of Remote Sensing*, Vol. 42, No. 7, 2021, pp. 2440–2462.

[15] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A., "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, Vol. 16, No. 1, 2024, pp. 45–74.

[16] licenta, "HERIDAL Dataset," `https://universe.roboflow.com/licenta-ynwvo/heridal-lrbkc`, jun 2022. URL `https://universe.roboflow.com/licenta-ynwvo/heridal-lrbkc`, visited on 2024-05-30.

[17] Hinniger, C., and Rüter, J., "Synthetic Training Data for Semantic Segmentation of the Environment from UAV Perspective," *Aerospace*, Vol. 10, No. 7, 2023. https://doi.org/10.3390/aerospace10070604.

[18] Rüter, J., and Schmidt, R., "Using Only Synthetic Images to Train a Drogue Detector for Aerial Refueling," *International Conference on Modelling and Simulation for Autonomous Systems (MESAS)*, 2023.

[19] European Aviation Safety Agency (EASA) and Deadalean AG, "Concepts of Design Assurance for Neural Networks (CoDANN)," Tech. rep., 2020.

[20] Federal Aviation Agency (FAA) and DaedaleanAG, "Neural Network Based Runway Landing Guidance for General Aviation Autoland," 2021.

[21] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Surface Vehicle Recommended Practice J3016," , 2016.

[22] The British Standards Institution, Center for Connected and Autonomous Vehicles, "PAS 1883:2021 Operational Design Domain (ODD) Taxonomy for an Automated Driving System (ADS) – Specification," , 2021. URL https://www.bsigroup.com/globalassets/localfiles/en-th/cav/bsi-cav-safety-benchmarking-report-2021-th.pdf.

[23] 33, I. S., "Road Vehicles – Test scenarios for automated driving systems – Specification for operational design domain," Standard, International Organization for Standardization, Aug. 2023.

[24] Torens, C., Juenger, F., Schirmer, S., Schopferer, S., Zhukov, D., and Dauer, J. C., *Ensuring Safety of Machine Learning Components Using Operational Design Domain*, AIAA, 2023. https://doi.org/10.2514/6.2023-1124, URL https://arc.aiaa.org/doi/abs/10.2514/6.2023-1124.

[25] Kaakai, F., Adibhatla, S. S., Pai, G., and Escorihuela, E., "Data-Centric Operational Design Domain Characterization for Machine Learning-Based Aeronautical Products," *Computer Safety, Reliability, and Security*, edited by J. Guiochet, S. Tonetta, and F. Bitsch, Springer Nature Switzerland, Cham, 2023, pp. 227–242. https://doi.org/https://doi.org/10.1007/978-3-031-40923-3_17, URL https://link.springer.com/chapter/10.1007/978-3-031-40923-3_17.

[26] Belcaid, M., Bonnafous, E., Crison, L., Faure, C., Jenn, E., and Pagetti, C., "Certified ML Object Detection for Surveillance Missions," , 2024. URL https://arxiv.org/abs/2406.12362.

[27] Lee, K., Lee, K., Lee, H., and Shin, J., "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," , 2018. URL https://arxiv.org/abs/1807.03888.

[28] Yang, J., Zhou, K., Li, Y., and Liu, Z., "Generalized Out-of-Distribution Detection: A Survey," *International Journal of Computer Vision*, Vol. 132, No. 12, 2024, pp. 5635–5662. https://doi.org/10.1007/s11263-024-02117-4, URL https://doi.org/10.1007/s11263-024-02117-4.

[29] Rüter, J., Maienschein, T., Schirmer, S., Schopferer, S., and Torens, C., "Filling the Gaps: Using Synthetic Low-Altitude Aerial Images to Increase Operational Design Domain Coverage," *Sensors*, Vol. 24, No. 4, 2024. https://doi.org/10.3390/s24041144, URL https://www.mdpi.com/1424-8220/24/4/1144.

[30] Guérin, J., Delmas, K., Ferreira, R. S., and Guiochet, J., "Out-Of-Distribution Detection Is Not All You Need," , 2023. URL https://arxiv.org/abs/2211.16158.

[31] Barredo Arrieta, A., DÃaz-RodrÃguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, Vol. 58, 2020, pp. 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.

[32] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al., "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, Vol. 106, 2024, p. 102301. https://doi.org/10.1016/j.inffus.2024.102301.

[33] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D., "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, Vol. 51, No. 5, 2018, pp. 1–42. https://doi.org/https://doi.org/10.1145/3236009.

[34] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C., "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, Vol. 16, No. none, 2022, pp. 1 – 85. https://doi.org/10.1214/21-SS133.

[35] Kiefer, B., Messmer, M., and Zell, A., "Diminishing domain bias by leveraging domain labels in object detection on UAVs," *20th International Conference on Advanced Robotics (ICAR)*, IEEE, 2021.

[36] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M., "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475. Used implementation: https://github.com/WongKinYiu/yolov7, accessed 22 June 2023.

[37] Peng, Y., "Athec," *Computational Communication Research*, Vol. 4, No. 1, 2022. https://doi.org/https://doi.org/10.5117/CCR2022.1.009.PENG, URL https://www.aup-online.com/content/journals/10.5117/CCR2022.1.009.PENG.

[38] Molnar, C., *Interpretable machine learning*, Lulu. com, 2020.

[39] Parodi, S., Filiberti, R., Marroni, P., Libener, R., Ivaldi, G., Mussap, M., Ferrari, E., Manneschi, C., Montani, E., and Muselli, M., "Differential diagnosis of pleural mesothelioma using Logic Learning Machine," *BMC bioinformatics*, Vol. 16 Suppl 9, 2015, p. S3. https://doi.org/10.1186/1471-2105-16-S9-S3.

[40] Muselli, M., "Switching Neural Networks: A New Connectionist Model for Classification," , 01 2005. https://doi.org/10.1007/11731177_4.

[41] Muselli, M., and Quarati, A., "Reconstructing positive Boolean functions with shadow clustering," *Proceedings of the 2005 European Conference on Circuit Theory and Design, 2005.*, Vol. 3, IEEE, 2005, pp. III–377.