

MULTIMODAL DATA FUSION FOR SECURITY
SCREENING WITH WEIGHTED CONFIDENCE
AGGREGATION AND ESCALATION IN EDGE-CLOUD
SYSTEMS

Serhii Robotko¹, Andrii Topalov^{1✉}, Igor Nechai²,
Roman Vasylenko³

Received on November 3, 2025

Presented by Ch. Roumenin, Member of BAS, on December 16, 2025

Abstract

Humanitarian demining requires rapid, accurate and resource-efficient identification of concealed explosive threats. This article presents a multimodal sensor-fusion framework that unites outputs from a visual detector (YOLOv8), two vision language models (ChatGPT-4.1 Vision and Gemini 2.5 Flash Vision) and a custom metal detector into a single integral probability P_{final} . Each modality's raw confidence is calibrated to a comparable probability via temperature or Platt scaling, then fused through learnable non-negative weights or logical rules (OR, AND, k -of- n) to support tunable trade-offs between probability of detection and false-alarm rate. A staged architecture executes preliminary detection on a UAV, confirms candidates at a ground station and escalates cases to the cloud, preserving real-time performance despite limited bandwidth and compute resources.

Key words: multi-level image analysis, multimodal data fusion, humanitarian demining, computer vision, YOLOv8, ChatGPT, cloud data processing

Introduction. Humanitarian demining requires fast, safe and low-false-alarm detection of buried threats. Vision-only pipelines struggle with occlusions, clutter and look-alike debris, while metal detectors alone trigger excessive alarms

from benign shrapnel and fragments. Recent progress in vision detectors and vision language models (VLM/LLMs) enables open-vocabulary reasoning over imagery, but principled fusion with non-visual sensors and deployment under edge limitations remain underexplored. In the development of UAV for humanitarian demining, in addition to improving visual detection systems, additional analytical models are also used. For example, approaches from [1], which developed an algorithm for predicting the sorption rate of hazardous gaseous substances, can be adapted for reconnaissance UAV equipped with chemical analysis sensors. Furthermore, mechanical deformations of rotor machine components [2] play an important role, which is taken into account in the development of more reliable unmanned aerial vehicles operating in challenging field conditions. Drone remote control systems also require special attention, as the effectiveness of the tasks performed depends on them [3–5].

We address this gap with a multimodal data-fusion framework that unifies YOLO-type visual detections, VLM/LLM semantic verification, and metal-detector measurements into a single integral probability P_{final} . Confidence scores from each source (s_Y, s_C, s_G, s_M) are calibrated and aggregated via learnable weights and rule-based decision policies (OR, AND, k -out-of- n), embedded in a staged escalation pipeline across edge, ground and cloud. The design allows sensitivity trade-offs to be tuned explicitly under bandwidth and compute budgets.

The model is instantiated on a UAV platform for land-mine identification, where the edge performs filtering, the ground node confirms candidates, and the cloud resolves hard cases. The paper includes quantitative charts, ROC and precision-recall curves, DET plots, calibration reliability diagrams, ablations over weights and thresholds, and latency bandwidth profiles—documenting gains in probability of detection at fixed false-alarm rates. To our knowledge, this is the first formal treatment of weighted confidence aggregation that jointly fuses visual detectors, VLM/LLM outputs, and metal-detector signals into a single decision variable for this class of systems.

Related work. Multisensory fusion has long been organized along the JDL model and the data/feature/decision taxonomy [6], with decision-level fusion using rules such as OR, AND and k -out-of- n . Classical approaches include Bayesian fusion, Dempster–Shafer evidence theory, weighted voting and learned weighting schemes; they provide complementary trade-offs between robustness to sensor faults and control over false alarms. Majority-logic and k -of- n [7] rules are popular in distributed detection due to simplicity and analyzable error bounds, while soft-decision variants improve sensitivity when communication is not the bottleneck.

In vision, modern one-stage detectors (e.g., YOLO family) deliver real-time proposals on embedded hardware, and VLM/LLMs enable open-vocabulary verification and context-aware disambiguation. Prior work has fused visual detectors with LiDAR or acoustic cues, magnetometers [8], but principled fusion with metal-

detector signals for humanitarian demining is sparse. Traditional mine-action literature emphasizes sensor complementarity and workflow triage. Our framework extends this line by adding semantic visual verification and an explicit calibrated probability P_{final} .

The study of LEKHAK et al. [9] released a UAV VNIR hyperspectral benchmark for landmine/UXO: 143 surrogates, Headwall Nano (270 bands, 398–1002 nm), with calibrated mosaics, raw cubes, GCPs, and high-fidelity spectra (RMSE < 1.0; spectral angle 1–6°). It fills a key gap and nicely complements our RGB+metal data for future multimodal fusion.

Overall, the proposed work combines established principles of decision-level fusion and score calibration with recent advances in open-vocabulary visual reasoning, yielding a practical, resource-aware framework for UAV-based land-mine identification and reporting.

Problem statement. This article tackles near real time land mine detection on a low power UAV that must work under strict limits on battery, computation and radio bandwidth while keeping recall high and false alarms low. Figure 1 schematically shows the drone with four sensing modules a visual detector that outputs a confidence score s_Y , a vision language or large language model that provides a semantic score s_C and optionally a grounding score s_G , and a metal detector whose signal is mapped to s_M . After calibration with temperature scaling and Platt scaling, these scores are fused with non negative trainable weights into a final probability P_{final} using policies such as weighted sum, OR or AND logic, or k -of- n voting. The threshold on P_{final} is tuned on a validation set to maximize detection probability and maintain precision under class imbalance while controlling false alarms, and decisions pass through three stages on board screening, ground station confirmation, and cloud review for ambiguous cases to remain robust under intermittent connectivity.

Mathematical model of weighted fusion. Modern sensor fusion integrates outputs from diverse detectors to produce more accurate and reliable decisions than any individual sensor can achieve. In our setup a vision language model (VLM) [10], an optional grounding network, and a metal detector each produce scores indicating the likelihood of a land mine. Data fusion combines these scores so that the final estimate benefits from the strengths of each modality; this approach is widely used in practice to increase detection reliability and reduce uncertainty [6].

To make the scores comparable we first calibrate each modality’s output into a probability in the $[0, 1]$ range. Visual detectors often output raw logits or confidence scores that are not well aligned with true likelihoods, and calibration techniques such as Platt scaling and temperature scaling fit a logistic regression model to map these scores into probabilities. Semantic verification and grounding scores from VLM/LLMs are similarly transformed, and if a sensor already delivers a probability, it is used directly. The result is a set of calibrated probabilities P_Y ,

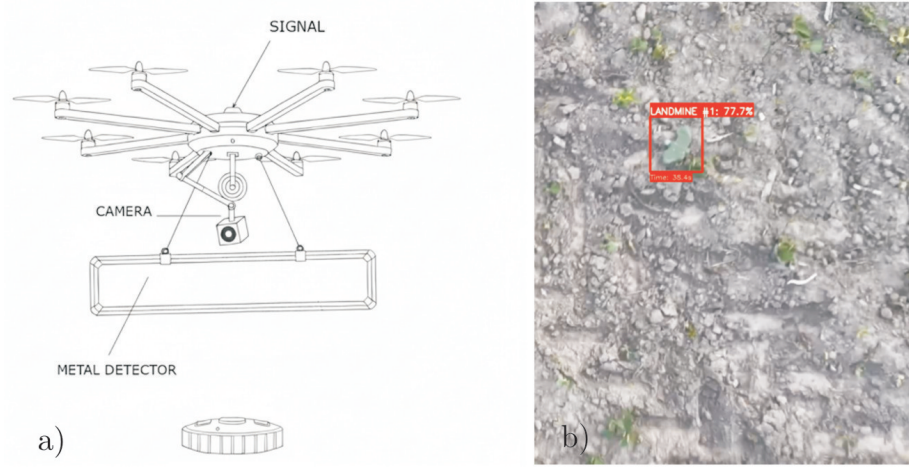


Fig. 1. Drone: a) in schematic view; b) object detection

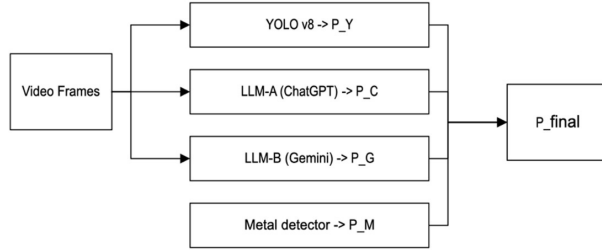


Fig. 2. Visual scheme of math model

P_C , P_G , P_M (Fig. 2) for vision, captioning, grounding, and metal detection, respectively. We then compute a single integral probability P_{final} as a weighted combination of the calibrated probabilities. A simple and interpretable form is a convex sum:

$$(1) \quad P_{\text{final}} = w_Y P_Y + w_C P_C + w_G P_G + w_M P_M,$$

where $w_i \geq 0$ and $w_Y + w_C + w_G + w_M = 1$. The weights reflect the relative trust placed on each modality; for example, a site with more metallic debris may use a lower w_M . Thresholding P_{final} at a value θ yields a binary decision: if $P_{\text{final}} \geq \theta$, a mine is declared, otherwise no mine is reported. Calibration and weight tuning on a validation set allow the user to set θ so that probability of detection and false-alarm rate meet operational requirements.

Alternative decision rules can be derived from the calibrated probabilities. The OR rule defines the probability that at least one modality detects a threat as:

$$(2) \quad P(D_{\text{final}} = 1 | E = 1) = 1 - (1 - P_Y)(1 - P_C)(1 - P_G)(1 - P_M),$$

which achieves high sensitivity but can increase false alarms. The AND rule multiplies all probabilities and is conservative, while k -out-of- n logic declares a detection if at least k of the n modality-specific probabilities exceed their thresholds [7]. These rules provide simple alternatives to the weighted sum when the priority is either maximum recall (OR) or maximum precision (AND). By selecting weights, calibration parameters and thresholds appropriately, the proposed fusion model balances detection probability and false-alarm rate and supports staged escalation across edge, ground and cloud nodes.

Experimental results. Field trials were carried out on an integrated aerial platform combining a quadrotor equipped with a Pixhawk 6C autopilot [11], a Raspberry Pi [12] camera and a custom-built metal detector. The system was tested in realistic demining conditions, with sources of noise such as grassy sur-

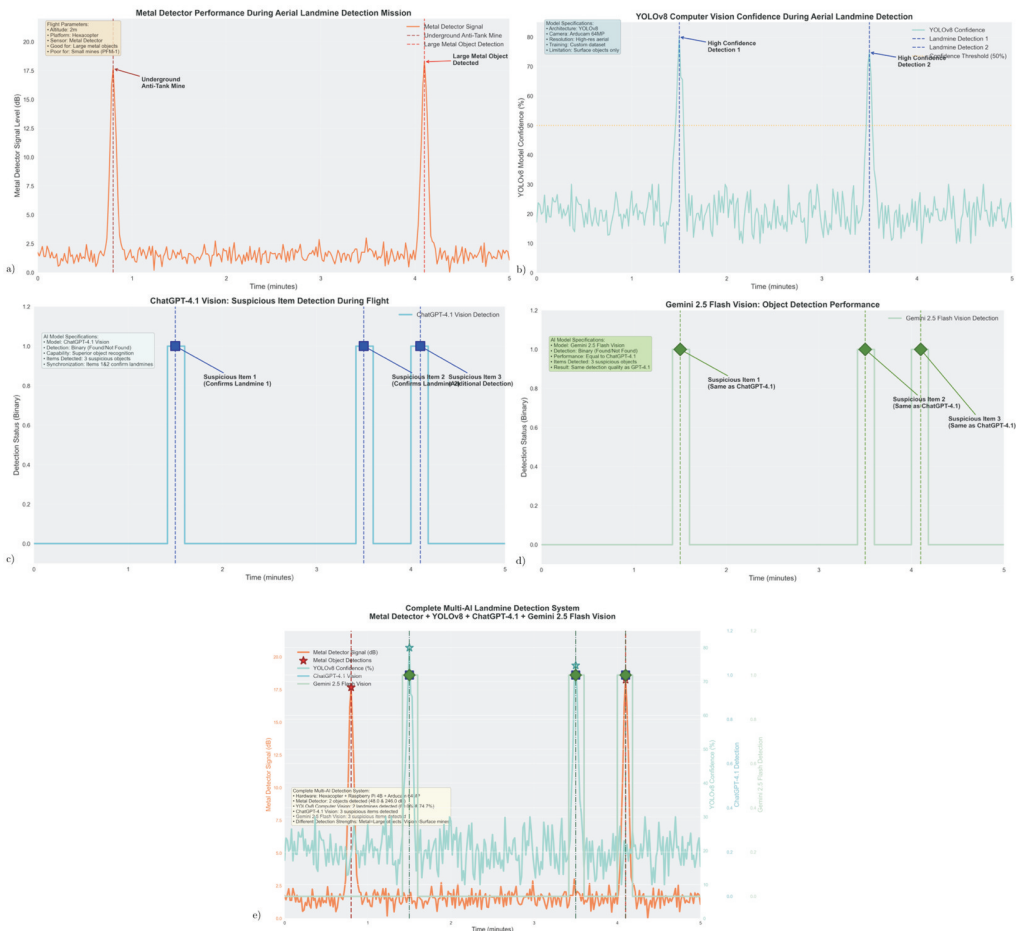


Fig. 3. Land-mine detecting by: a) metal detector; b) YOLOv8 model; c) ChatGPT-4.1 Vision; d) Gemini 2.5 Flash Vision; e) complete chart

faces, soil heterogeneity, shadows and micro-vibration. Each flight lasted approximately 15 minutes to collect enough data. The altitude was held between 2 and 3 m using LiDAR to keep a stable object scale and safe clearance. During flights, the Raspberry Pi recorded 1920×1080 video (30fps) and digitized the metal-detector audio. Both channels were timestamped and synchronized using network time protocol (NTP). After landing, the videos and logs were sent via Wi-Fi/SSH to a laptop where pre-processing and inference were run, with georeferencing and annotation performed post-hoc. Metal-detector (Fig. 1a) calibration on clean patches of ground established baseline thresholds. Zero signal did not guarantee absence of a buried object (because of small plastic land-mines), whereas a positive signal indicated a strong subsurface anomaly.

The computer-vision module was evaluated separately to quantify its performance and limitations. We trained YOLOv8n from scratch (no pretraining) on 650 manually labelled UAV frames of PFM-1 mines captured at low altitude under varied angles, lighting, and vibration-induced blur. Training used SGD ($\text{lr} = 0.01$) for up to 50 epochs with early stopping (patience = 20) and batch size was 8 on GPU/MPS. Checkpoints were saved every 10 epochs, and the best model was chosen by performance on a held-out validation set. A validation set contained 53 images and 23 instances of PFM1 mines showing three scene types: open soil/grass, partial occlusion by vegetation and fragments with heterogeneous backgrounds. Performance was assessed using standard metrics (precision, recall, $\text{mAP}@0.5$ and $\text{mAP}@0.5 = 0.95$) and productivity metrics (per-frame processing time and effective frame rate).

The YOLOv8n (fused) model [13] achieved a precision of 0.957, recall of 1.000 and $\text{mAP}@0.5$ of 0.981. The more stringent $\text{mAP}@0.5 = 0.95$ dropped to 0.688 because stricter intersection-over-union (IoU) thresholds penalize small scale changes and partial occlusions. Inference times averaged 0.5 ms for pre-processing, 21.2 ms for inference and 6.0 ms for post-processing, giving roughly 27.7 ms per frame (≈ 36 FPS) on an Apple M3-Pro laptop. Most latency was spent in inference ($\sim 76.5\%$), suggesting optimization opportunities via lower-resolution inputs, quantization or hardware accelerators.

Table 1 summarizes detections from four subsystems-metal detector, YOLOv8, ChatGPT-4.1 Vision and Gemini 2.5 Flash Vision-over the same 15-minute flight.

The metal detector produced two strong amplitude peaks (17.7=18.3 dB) corresponding to buried metallic anomalies at 0.8 and 4.1 minutes. The YOLOv8 model detected two surface-visible mines at about 1.5 and 3.5 minutes with high confidence. Its confidence dropped at other times, indicating no detections. ChatGPT-4.1 Vision and Gemini 2.5 Flash Vision operate as semantic verifiers – they take cropped image patches around YOLO bounding boxes and output binary flags indicating whether the patch depicts an explosive object. These models flagged three suspicious episodes, confirming the two high-confidence YOLO

T a b l e 1
Detection summary

AI Model	Metal detector	YOLOv8	ChatGPT-4.1 Vision	Gemini 2.5 Flash Vision
Detection type	Hardware signal	Confidence %	Binary detection	Binary detection
Objects detected	2	2	3	3
Detection times (min)	0.8, 4.1	1.5, 3.5	1.5, 3.5, 4.1	1.5, 3.5, 4.1
Performance	17.7–18.3 dB	80.0–74.7%	3 suspicious items	3 suspicious items

detections and adding a third event at 4.1 minutes that coincided with a metal-detector alarm but had lower YOLO confidence. This additional trigger highlights the value of multimodal verification for ambiguous frames.

To characterize temporal behaviour of the vision detector, a confidence-tracing graph (Fig. 2b) plots the YOLOv8 confidence over flight time. The graph shows pronounced peaks around 1.5 and 3.5 minutes at $\sim 80.0\%$ and $\sim 74.7\%$ confidence. Annotated frames at these times show clear PFM-1 silhouettes with well-defined contours and consistent aspect ratios. At around 0.8 minutes there is no response, demonstrating a physical limitation of the optical channel because when targets are fully occluded by soil or vegetation, the purely visual model justifiably returns low scores. Figure 3b provides a qualitative example of a surface detection. In a 1920×1080 frame, YOLOv8 returned a single bounding box labelled LANDMINE with 77.7% confidence. The object was localized above the working IoU threshold and passed non-maximum suppression. The mine lay on loose soil with small vegetation, and the bounding box aligned closely with the visual silhouette.

The ChatGPT-4.1 Vision [14] model was integrated as a post-hoc semantic verifier: it receives cropped image patches and captions them using an open-vocabulary multimodal model. The resulting descriptive sentences are compared against a set of keywords (e.g., “landmine”, “mine casing”) to produce binary flags. A graph (Fig. 2c) of ChatGPT-4.1’s output over time shows binary spikes at the same 1.5 and 3.5 minute marks as the YOLO [15] confidence peaks, confirming those detections. It also flags a third object around 4.1 minutes. Preliminary experiments estimated the accuracy of ChatGPT-4.1 in this classification role at approximately 64% ($P_C = 0.64$), reflecting that the model is designed for rich description rather than precise detection

A second vision-language model was Gemini 2.5 Flash [16] evaluated for the same task. Its graph (Fig. 2d) shows binary triggers aligned with the ChatGPT-4.1 peaks, including a notable episode at ~ 4.1 minutes, which provides inter-model agreement. Both models occasionally trigger early in the mission around ~ 0.8

minutes. These spurious activations emphasize fundamental limitations of purely visual methods for detecting subsurface objects. Experimental analysis estimated the accuracy of Gemini 2.5 Flash at roughly 68% ($P_G = 0.68$).

The final chart (Fig. 2e) overlays outputs from all four detection systems—metal detector, YOLOv8 confidence, ChatGPT-4.1 Vision, and Gemini 2.5 Flash Vision over the 15-minute flight. Two strong metal-detector signals at 0.8 and 4.1 minutes indicate buried metallic anomalies. The YOLOv8 confidence trace peaks at 1.5 and 3.5 minutes, corresponding to surface-visible mines. The language-model triggers align with the YOLO peaks and add an extra flag at 4.1 minutes. This combined view demonstrates the complementarity of modalities as surface mines are captured reliably by the vision detector, buried anomalies are indicated by the metal detector [17], and ambiguous cases are highlighted by language-model verification. Such complementary evidence motivates the weighted-fusion approach presented earlier.

To calculate the overall detection probability P_{detect} under the OR-rule fusion, we use the complement rule from probability theory. This rule states that the probability of at least one of several independent events occurring is equal to 1 minus the probability that none of the events occur. In our context the events are detections by the metal detector (P_M), the visual detector (P_Y), ChatGPT-4.1 Vision (P_C), and Gemini 2.5 Flash Vision (P_G). The formula is:

$$(3) \quad P_{\text{detect}} = 1 - (1 - P_Y) (1 - P_C) (1 - P_G) (1 - P_M).$$

Substituting the given probabilities $P_M = 0.65$, $P_Y = 0.80$, $P_C = 0.64$ and $P_G = 0.68$ yields:

$$(4) \quad P_{\text{detect}} = 1 - (1 - 0.80) (1 - 0.64) (1 - 0.68) (1 - 0.65) = 0.9919.$$

When using an OR decision rule, the probability that at least one of the four subsystems detects a threat is approximately 99.1%. This high value illustrates how fusing independent modalities dramatically increases the likelihood of detection compared with relying on any single sensor.

Conclusions. We presented a resource aware multimodal framework for land mine detection on unmanned aerial vehicles that calibrates and fuses confidence scores from a visual detector, vision–language models and a metal detector into a single probability P_{final} . An edge-ground-cloud pipeline provides real time alerts on low power platforms while allowing more computationally intensive verification in the cloud, and tunable weights and thresholds let operators balance recall against false alarms.

In 15 minute flights over realistic terrain, the vision module alone reached perfect recall and high precision on PFM-1 targets, while multimodal OR fusion increased the overall detection probability to 99.1%. Semantic verifiers (ChatGPT 4.1 Vision and Gemini 2.5 Flash Vision) helped confirm detections and flag

ambiguous cases. The framework runs at about 36 FPS on consumer hardware, can absorb new sensors through the same calibration–fusion process, and will be extended to more mine types, terrains, adaptive weighting, and additional modalities such as compact semantic models and ground penetrating radar.

REFERENCES

- [1] LESKO A., O. KULAKOV, A. MELNICHENKO, A. KATUNIN (2024) Development and implementation of an algorithm for predicting the intensity of sorption of hazardous gaseous materials, *Solid State Phenomena*, **364**, 101–112, <https://doi.org/10.4028/p-qjra2j>.
- [2] HUI W., S. SHEVCHENKO, M. KORCHAK et al. (2025) A study of mechanical seal rings thermal and force deformations in energy rotor machines, *C. R. Acad. Bulg. Sci.*, **78**(1), 109–118, <https://doi.org/10.7546/CRABS.2025.01.13>.
- [3] GERASIN O. S., A. M. TOPALOV, M. O. TARANOV et al. (2020) Remote IoT-based control system of the mobile caterpillar robot, *CEUR Workshop Proceedings*, **2740**, 129–136.
- [4] KONDRATENKO Y., O. GERASIN, O. KOZLOV et al. (2021) Inspection mobile robot’s control system with remote IoT-based data transmission, *Journal of Mobile Multimedia*, **17**(4), 499–526, <https://doi.org/10.13052/jmm1550-4646.1742>.
- [5] KOZLOV O., Y. KONDRATENKO, O. SKAKODUB et al. (2023) Swarm Optimization of fuzzy systems for mobile robots with remote control, *Journal of Mobile Multimedia*, **19**(3), 839–876, <https://doi.org/10.13052/jmm1550-4646.1939>.
- [6] CASTANEDO F. (2013) A review of data fusion techniques, *Scientific World Journal*, **2013**(1), <https://doi.org/10.1155/2013/704504>.
- [7] ZHU J., H. CAO, J. MU (2019) A further exploration of multi-slot based spectrum sensing, *Sensors (Basel)*, **19**(6), 3497, <https://doi.org/10.3390/s19163497>.
- [8] STANKEVICH S. A., I. Y. SAPRYKIN (2024) Optical and magnetometric data integration for landmine detection with UAV, *WSEAS Transactions on Environment and Development*, **20**, <https://doi.org/10.37394/232015.2024.20.96>.
- [9] LEKHAK S., E. J. IENTILUCCI, J. BAUR, S. GHOSH (2025) A UAV-based VNIR hyperspectral benchmark dataset for landmine and UXO detection, *Electrical Engineering and Systems Science*, <https://arxiv.org/abs/2510.02700>.
- [10] ROBOTKO S. (2025) AI-driven OCR in practice: assessing ChatGPT-4.1 and Google Vision as alternatives to manual optical character recognition, *15th International Conference on Advanced Computer Information Technologies (ACIT)*, <https://doi.org/10.1109/ACIT65614.2025.11185827>.
- [11] ArduPilot Documentation (2024) <https://ardupilot.org/ardupilot/>.
- [12] MONK S. (2021) *Programming the Raspberry Pi. Getting Started with Python*, 3rd Edn, McGraw Hill TAB, 104–220.
- [13] Ultralytics YOLO Docs, <https://docs.ultralytics.com>.
- [14] ROTHMAN D. (2024) *Transformers for Natural Language Processing and Computer Vision: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3*, Packt Publishing, 88–144.

- [15] LI F., Y. ZHANG, Y. FAN (2025) EABI-DETR: An efficient aerial small object detection network, *Biomimetics*, **10**(11), 770, <https://doi.org/10.3390/biomimetics10110770>.
- [16] Gemini 2.5 Flash Documentation, <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash>.
- [17] PERRY (2024) Saving Lives, Step by Step: Robust Techniques for Landmine Detection, *Tredition GmbH*, 191–243.

¹*Department of Computerized Control Systems,
Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine
e-mails: robotkos@gmail.com, topalov_ua@ukr.net*

²*National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine
e-mail: i.nechai@nubip.edu.ua*

³*Ivan Kozhedub Kharkiv National Air Forces University, Kharkiv, Ukraine
e-mail: spike75.rv@gmail.com*