Missing-Modality-Tolerant Feature-Level Fusion for Drone and Aerial Target Classification

Luis Paulo Albuquerque Guedes[1], Marco André Desbrousses Cotta[2] e Danilo Olivieri[3]
[1]Comando do 7º Distrito Naval (Com7DN), Brasília/DF - Brasil
[2]Centro de Guerra Acústica e Eletrônica da Marinha (CGAEM), Rio de Janeiro/RJ - Brasil
[3]Instituto de Pesquisas da Marinha (IPqM), Rio de Janeiro/RJ - Brasil

***Abstract*— This work investigates missing-modality-tolerant multimodal classification of aerial targets using infrared (IR), visible (V), and acoustic (AUDIO) sensors. Using the public *Drone Detection Thesis Dataset*, seven unimodal and multimodal sensing scenarios were evaluated through record-level feature representations. Missing sensors were handled by feature imputation and explicit modality-presence indicators. Classical classifiers with early fusion, late fusion, weighted fusion, and stacking were compared with deep architectures over ten independent repetitions. The best result was obtained with IR+V+AUDIO early fusion and *HistGradientBoosting*, achieving balanced accuracy of 0.9005, macro-F1 of 0.9002, and weighted-F1 of 0.9082. AUDIO was the strongest unimodal source, while the evaluated classical models outperformed the deep architectures. These results indicate that feature-level fusion with explicit missing-modality handling is effective for small, heterogeneous multisensor datasets.**

***Keywords*— Drones, multimodal fusion, machine learning.**

I. Introduction

Drone detection and classification are relevant to defense, public security, and critical infrastructure protection. Their automatic identification remains challenging due to small target size, similarity to birds, and environmental variability. Visible, infrared, and acoustic sensors provide complementary information but are affected by illumination, thermal contrast, noise, distance, and atmospheric conditions [1]–[5]. Radar and RF approaches face additional limitations related to low radar cross-section, clutter, autonomous operation, and encrypted links [4]–[6]. These limitations motivate multimodal fusion.

This work uses the public *Drone Detection Thesis Dataset*, containing infrared and visible videos and audio clips from the classes *airplane*, *bird*, *drone*, *helicopter*, and *background* [3]. Unlike the original YOLOv2 and MFCC+LSTM approach, we evaluate feature-based multiclass classification using IR, V, and AUDIO data.

Seven sensing scenarios are evaluated: IR, V, AUDIO, IR+V, IR+AUDIO, V+AUDIO, and IR+V+AUDIO. Missing modalities are handled through feature imputation and sensor-presence indicators. Classical classifiers, early and late fusion, weighted fusion, stacking, and deep models are compared. Since modality indicators may encode dataset-availability patterns, their isolated effect is considered a limitation and left for future ablation studies. The main contribution is a comparative evaluation of missing-modality-tolerant multimodal fusion for multiclass aerial target classification on a public multisensor dataset.

L. P. A. Guedes, luis.guedes@coppe.ufrj.br; M. A. D. Cotta, macotta@poli.ufrj.br; D. Olivieri, danilo.olivieri@marinha.mil.br.

II. Related Work

Automatic drone detection has been studied using optical, thermal, acoustic, radar, radio-frequency, and multisensor approaches. Recent surveys highlight its relevance to airport security, critical infrastructure protection, and surveillance of sensitive areas [7]–[9]. The main challenges include small target size, similarity to birds, distance variation, limited public datasets, and integration of heterogeneous sensors.

In the visual domain, convolutional networks and object detectors have been applied to drone localization in images and videos [10]–[13]. Other studies address drone–bird discrimination to reduce false alarms in open environments [14], [15]. Despite good performance under favorable conditions, visual methods remain sensitive to illumination, distance, occlusion, and target similarity.

Acoustic sensors exploit the sound signature of motors and rotors, providing a complementary alternative to cameras. Previous studies investigated acoustic identification using handcrafted features, deep neural networks, and data augmentation techniques [16]–[19]. However, acoustic performance may degrade under ambient noise, distance, wind, and interference from other sound sources.

Sensor fusion is a promising strategy to improve the robustness of counter-drone systems [8], [9]. Previous works combined acoustic and visual information or proposed multimodal deep learning frameworks [20]–[22]. These approaches exploit the complementarity among modalities: cameras capture shape and motion, thermal sensors exploit infrared contrast, and acoustic sensors represent rotor sound patterns.

The work by Svanström et al. [23] is particularly relevant because it proposes a real-time system using visible, thermal, and acoustic sensors, while also releasing a public multisensor dataset with infrared videos, visible videos, and audio clips [24]. More recent datasets, such as MMAUD [25], extend this direction by incorporating multiple modalities, including stereo vision, LiDAR, radar, and acoustic arrays.

Despite these advances, many studies prioritize visual detection, isolated acoustic classification, or decision-level fusion. Systematic evaluations of multiclass classification with missing modalities are still less common in public datasets containing infrared, visible, and acoustic sensors. Therefore, this work complements the literature by comparing early fusion, late fusion, and stacking across seven sensing scenarios: IR, V, AUDIO, IR+V, IR+AUDIO, V+AUDIO, and IR+V+AUDIO.

III. Materials and Methods

This section describes the dataset, sensing scenarios, feature representation, classical and deep models, fusion strategies, and experimental protocol. Fig. 1 summarizes the overall architecture of the implemented pipeline.

A. Dataset

The experiments were conducted using the public *Drone Detection Thesis Dataset*, available at `https://github.com/DroneDetectionThesis/Drone-detection-dataset`. This dataset was originally proposed for drone detection and tracking using visible, thermal, and acoustic sensors [23], [24].

The dataset contains thermal infrared videos, visible-spectrum videos, and audio clips. The video modalities include the classes *airplane*, *bird*, *drone*, and *helicopter*, while the acoustic modality includes *drone*, *helicopter*, and *background*. In this work, these labels were mapped to the global class space AIRPLANE, BACKGROUND, BIRD, DRONE, and HELICOPTER.

B. Proposed Pipeline

Fig. 1 presents the general workflow. First, tabular feature files from the IR, V, and AUDIO sensors are loaded and standardized. Then, unimodal and multimodal scenarios are constructed. For classical models, unimodal classification, early fusion, late fusion by averaging, weighted late fusion, and stacking are evaluated. For deep models, evaluation is restricted to the IR+V+AUDIO scenario using early feature fusion.

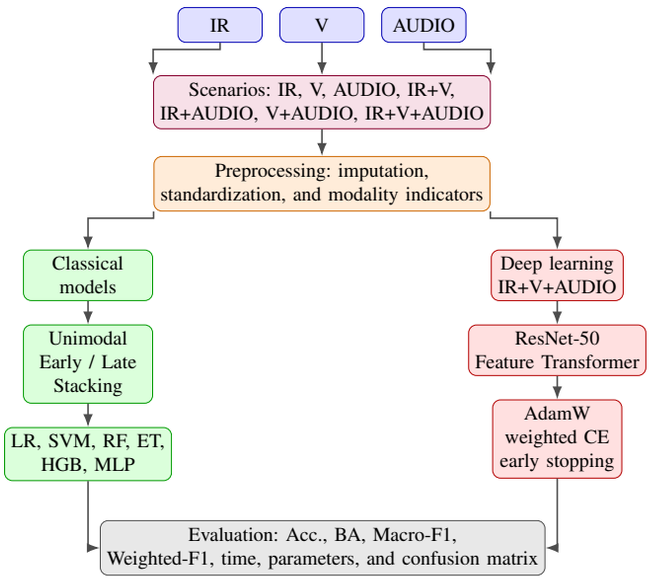

Figure 1. Compact pipeline for multiclass classification using classical and *deep learning* models.

C. Sensing Scenarios

Seven experimental scenarios were evaluated, as shown in Table I. Classical models were evaluated in all scenarios, whereas deep models were evaluated only in the IR+V+AUDIO scenario, since it is the most complete multimodal setting.

TABLE I
EVALUATED SENSING SCENARIOS

Scenario	Modalities	Evaluation
IR	Infrared	Classical
V	Visible	Classical
AUDIO	Audio	Classical
IR+V	Infrared + Visible	Classical
IR+AUDIO	Infrared + Audio	Classical
V+AUDIO	Visible + Audio	Classical
IR+V+AUDIO	IR + Visible + Audio	Classical and Deep

Because not all classes are available in all modalities, a sample-union strategy was adopted. Thus, a sample is included in a multimodal scenario if it is present in at least one of its modalities. When a modality is unavailable, its features remain missing and are handled by imputation. In addition, explicit sensor-presence indicators, namely `has_IR`, `has_V`, and `has_AUDIO`, were added to distinguish an unavailable sensor from a valid observed feature vector. Since modality availability is not uniformly distributed across all classes in the dataset, these indicators may also encode dataset-availability patterns. Their isolated contribution is not evaluated in the present work and is therefore considered a limitation to be addressed through dedicated ablation studies in future work.

Therefore, the proposed formulation should be interpreted as a missing-modality-tolerant feature-level fusion protocol over heterogeneous sensor-derived samples, rather than as a strictly synchronized multisensor fusion setting.

D. Feature Representation

The experiments used three pre-extracted tabular feature files: `features_IR.csv`, `features_V.csv`, and `features_AUDIO.csv`. Each file contains one fixed-dimensional representation per dataset record, identified by `sample_key`, together with the class label `class_name` and the numerical attributes associated with the corresponding sensor. Consequently, the classifiers evaluated in this study operate on record-level feature vectors rather than on individual video frames or raw audio samples.

The IR file contains 365 samples and 2124 features, covering AIRPLANE, BIRD, DRONE, and HELICOPTER. The visible file contains 285 samples and 2124 features for the same four classes. The audio file contains 90 samples and 178 features, covering BACKGROUND, DRONE, and HELICOPTER. No missing feature values were observed within the original unimodal files.

For the IR and visible modalities, each complete video record is represented by a single fixed-length vector with 2124 numerical attributes. For the acoustic modality, each audio clip is represented by a single fixed-length vector with 178 numerical attributes. These feature vectors constitute the direct inputs to the classification and fusion stages analyzed in this work; therefore, no frame-level sampling is performed by the classifiers described here. The present study focuses on the behavior of the classification and multimodal-fusion strategies over these pre-extracted sensor representations. A more detailed reconstruction and evaluation of the raw-signal-to-feature extraction stage is outside the scope of the present experiments and is identified as an important reproducibility extension for future work.

Feature names were standardized by sensor prefix, resulting in columns such as `IR_f*`, `V_f*`, and `AUDIO_f*`. In multimodal scenarios, missing values arise only from the sample-union procedure, since not all samples and classes are available for every modality. The final input vector was formed by concatenating the available sensor features and explicit modality-presence indicators.

E. Preprocessing

For classical models, preprocessing was implemented through pipelines with median imputation and, when required, mean–standard deviation standardization. Standardization was applied to scale-sensitive models, such as logistic regression, support vector machines, and multilayer perceptrons. For tree-based models, standardization was not used.

For deep models, features were also median-imputed and standardized. The vectors were then reshaped according to the architecture. In the ResNet-50, the tabular vector was zero-padded to form a single-channel square map. In the *Feature Transformer*, the vector was split into fixed-size tokens.

F. Classical Models

The evaluated classical models are shown in Table II. The selection includes linear classifiers, tree-based models, ensemble methods, and a shallow neural network. Optional XGBoost, LightGBM, and CatBoost models were included in the code, but remained disabled in this experimental configuration.

TABLE II
EVALUATED CLASSICAL MODELS

Model	Family
Dummy prior	Baseline
Logistic Regression L2	Linear
Linear SVM	Maximum margin
Random Forest	Tree ensemble
Extra Trees	Randomized tree ensemble
HistGradientBoosting	Gradient boosting
MLPClassifier	Shallow neural network

All models were trained using the local labels of each evaluated scenario. After prediction, probabilities were mapped to the global class space, enabling comparison across scenarios with different class sets.

G. Classical Fusion Strategies

In multimodal scenarios, four main fusion strategies were evaluated, as summarized in Table III. Early fusion directly concatenates sensor features, including missing values and modality indicators. Late fusion combines probabilities estimated by classifiers trained separately for each sensor. Stacking uses these probabilities as input to a logistic meta-classifier.

The late-fusion strategies combine probabilistic outputs from classifiers trained separately for each modality. First, unimodal classifiers produce probability distributions in the global class space, namely $p_{\mathrm{IR}}(y|x)$, $p_{\mathrm{V}}(y|x)$, and $p_{\mathrm{AUDIO}}(y|x)$. These probabilities are then integrated using simple averaging, out-of-fold (OOF) performance-weighted averaging, or stacking with a meta-classifier. The final multi-class prediction is obtained from the aggregated output.

TABLE III
EVALUATED FUSION STRATEGIES

Strategy	Description
Unimodal	Sensor-specific training
Early concat missing	Concatenation with missing modalities
Late mean	Simple probability averaging
Late weighted OOF	OOF-performance-weighted averaging
Late stacking	Logistic meta-classifier

H. Deep Learning Models

In addition to the classical models, two deep architectures were evaluated in the IR+V+AUDIO scenario. The first was a ResNet-50 adapted to feature maps, in which the multimodal tabular vector is reorganized as a single-channel two-dimensional map. The second was a *Feature Transformer*, which represents the attributes as a sequence of tabular tokens. These architectures allowed classical models and neural approaches to be compared under the same experimental protocol.

Fig. 2 shows the organization of the two deep models. ResNet-50 receives the tabular vector converted into a two-dimensional map, while the *Feature Transformer* receives a sequence of tokens extracted from the feature vector.

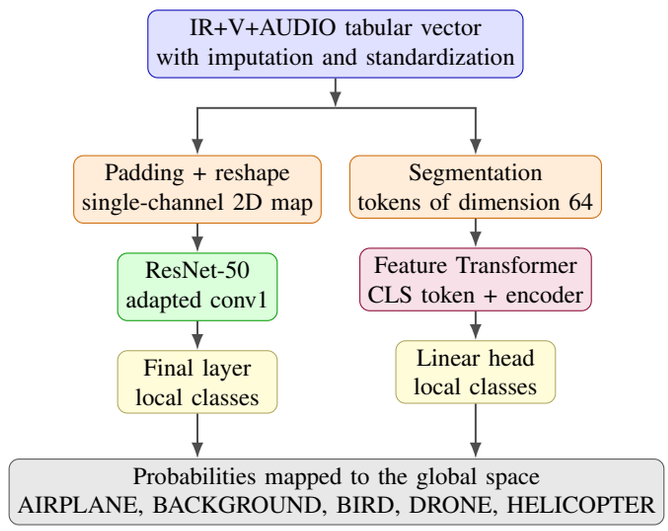

Figure 2. Representation of the deep models evaluated in the IR+V+AUDIO scenario.

In the ResNet-50 feature-map model, the initial convolutional layer was modified to receive a single-channel input. The final layer was replaced to output the number of local classes in the scenario. The network weights were initialized from scratch, without pretraining. The feature vector was zero-padded to form a square map of size $\lceil\sqrt{d}\rceil \times \lceil\sqrt{d}\rceil$, where d is the number of attributes.

In the *Feature Transformer*, the feature vector was segmented into tokens of dimension 64. Each token was projected to a latent dimension of 128, followed by a classification token, learnable positional encoding, and three Transformer encoder layers. The model used four attention heads, an intermediate dimension of 256, and a dropout rate of 0.15.

I. Training of Deep Models

The deep models were trained with weighted cross-entropy to account for class imbalance. The optimizer was AdamW,

with a learning rate of 3×10^{-4} and weight decay of 10^{-4}. Cosine learning-rate scheduling, gradient clipping with maximum norm 3.0, and automatic mixed precision were also used when a GPU was available.

Training was limited to 50 epochs, with early stopping after 10 epochs without improvement in the validation metric. For each repetition, the training set was split into internal training and validation subsets, using 20% of the training data for validation whenever stratification was possible. The monitored metric was validation macro-F1.

J. Experimental Protocol

The protocol used ten independent repetitions with different random seeds. In each repetition, record-level feature vectors (`sample_key`) were stratified into 70% training and 30% testing sets. Thus, frames from the same video were not independently split across sets.

Classical models were evaluated using unimodal inputs, early fusion, late fusion, weighted late fusion, and stacking. Deep models were evaluated only for IR+V+AUDIO using early fusion with missing-modality handling.

K. Metrics and Computational Effort

Performance was assessed using accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, and weighted-F1, with macro-F1 as the main metric over classes present in each test set.

Training and inference time, model size, deep-model parameter count, training epochs, and memory usage were also recorded. Per-repetition results, statistical summaries, predictions, leaderboards, and confusion matrices were saved.

IV. Results and Discussion

This section presents the results for the IR, V, AUDIO, IR+V, IR+AUDIO, V+AUDIO, and IR+V+AUDIO scenarios. The comparison uses balanced accuracy (BA), macro-F1, and weighted-F1. Since not all sensors contain all global classes, macro-F1 was computed only over the classes present in the test set of each scenario; global macro-F1 was kept only for auditing.

Since different sensing scenarios contain different subsets of the global label space, cross-scenario comparisons should be interpreted as performance under each available-label condition rather than as strictly identical classification tasks.

A. Evaluation Metrics

Accuracy may favor majority classes in imbalanced datasets. Therefore, balanced accuracy was also adopted, defined as the average recall across classes. The F1-score combines precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{1}$$

In multiclass problems, macro-F1 is computed as:

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^{C} F1_c, \tag{2}$$

where C is the number of evaluated classes. This metric was used as the main criterion because it assigns equal weight to each class present in the test set.

B. Best Results by Scenario

Table IV presents the best result for each scenario. The best overall performance was obtained in IR+V+AUDIO with early fusion and *HistGradientBoosting*, reaching BA of 0.9005 and macro-F1 of 0.9002.

Fig. 3 shows the aggregated confusion matrix of the best model. The BACKGROUND and DRONE classes were classified without errors, with 90 and 470 correct predictions, respectively. The main confusions occurred among BIRD, AIRPLANE, and HELICOPTER, although the largest values remained on the main diagonal.

The confusion matrix confirms strong overall performance for the IR+V+AUDIO *HistGradientBoosting* model, with most predictions on the main diagonal. BACKGROUND and DRONE were classified without errors, while the main confusions occurred among BIRD, AIRPLANE, and HELICOPTER. Under the available-label evaluation, IR+V+AUDIO achieved higher macro-F1 than the unimodal scenarios, with gains of 0.0698 over IR, 0.1689 over V, and 0.0192 over AUDIO.

C. Best Results by Model

Table V summarizes the best result obtained by each model. *HistGradientBoosting* achieved the best overall performance. *Late stacking* was competitive, but more computationally costly. *Extra Trees* achieved close performance with much faster training. The deep models remained below the best classical models.

The best classical model achieved macro-F1 of 0.9002, while the best deep model reached 0.8236. This gap suggests that, for the evaluated small tabular dataset, classical models were more suitable than deep architectures trained from scratch.

D. Computational Effort

Table VI presents training time, inference time, number of parameters, and model size for the main models. For deep models, the epoch column indicates the average number of executed epochs and the average best validation epoch.

HistGradientBoosting achieved the best overall performance with moderate cost: 32.96 s of training and 3.95 ms per sample. *Extra Trees* was the most efficient alternative in training, with macro-F1 of 0.8922 and only 1.32 s. *Late stacking* was competitive, but had higher cost due to the generation of out-of-fold predictions. Among deep models, ResNet-50 was heavier, while the *Feature Transformer* achieved the lowest latency, but with lower predictive performance.

E. Comparison with the Reference Work

Table VII compares the results with the reference work [23]. The comparison is not direct, since the reference evaluates detection with YOLOv2, acoustic classification with MFCC+LSTM, and confidence-based fusion, whereas this work evaluates multiclass classification using tabular features.

The values in Table VII provide contextual reference points rather than direct baselines, since the tasks, inputs, and metrics differ. The proposed protocol achieved competitive results, especially in the complete IR+V+AUDIO scenario, although its macro-F1 is not directly comparable to the 78% detection-opportunity rate reported in the reference.

TABLE IV
BEST RESULTS BY SCENARIO

Scenario	Family	Mode	Model	Accuracy	BA	Macro-F1	Weighted-F1
IR	Classical	Unimodal IR	HistGradientBoosting	0.8655 ± 0.0175	0.8299 ± 0.0247	0.8304 ± 0.0228	0.8632 ± 0.0166
V	Classical	Unimodal V	HistGradientBoosting	0.7860 ± 0.0511	0.7312 ± 0.0650	0.7313 ± 0.0676	0.7809 ± 0.0540
AUDIO	Classical	Unimodal AUDIO	Extra Trees	0.8815 ± 0.0483	0.8815 ± 0.0483	0.8810 ± 0.0490	0.8810 ± 0.0490
IR+V	Classical	Late stacking	LogReg meta	0.9107 ± 0.0109	0.8907 ± 0.0140	0.8890 ± 0.0145	0.9104 ± 0.0110
IR+AUDIO	Classical	Unimodal AUDIO	Linear SVM	0.8869 ± 0.0385	0.8878 ± 0.0420	0.8888 ± 0.0397	0.8867 ± 0.0386
V+AUDIO	Classical	Unimodal AUDIO	Logistic Regression L2	0.8826 ± 0.0500	0.8793 ± 0.0526	0.8760 ± 0.0545	0.8773 ± 0.0542
IR+V+AUDIO	Classical	Early concat missing	HistGradientBoosting	$\mathbf{0.9091 \pm 0.0255}$	$\mathbf{0.9005 \pm 0.0284}$	$\mathbf{0.9002 \pm 0.0277}$	$\mathbf{0.9082 \pm 0.0258}$

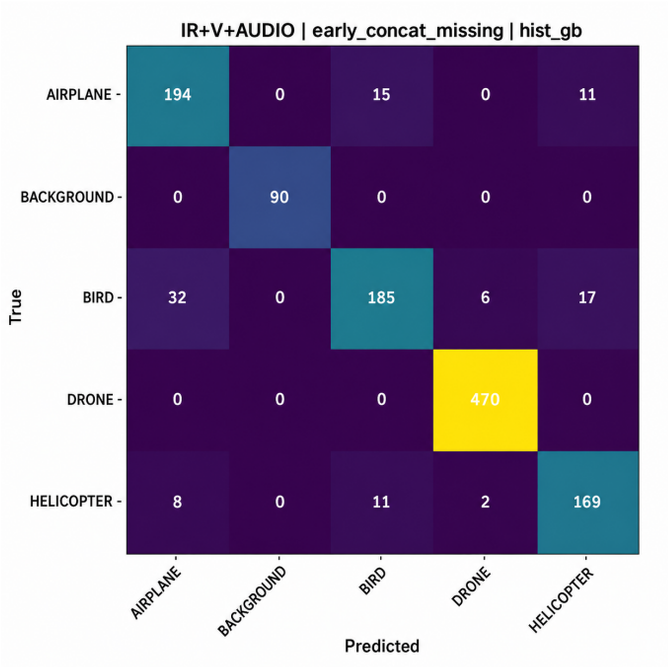

Figure 3. Aggregated confusion matrix of the best overall model, corresponding to the IR+V+AUDIO scenario with early fusion and the *HistGradientBoosting* classifier.

F. Final Analysis

The results indicate that IR+V+AUDIO provided the best overall performance, with *HistGradientBoosting* offering the best performance–cost trade-off. AUDIO was the strongest unimodal source, while classical models outperformed the evaluated deep architectures. *Extra Trees* remains an efficient alternative when training time is critical.

Since the sensors cover different class subsets, the scenarios should be interpreted as missing-modality-tolerant feature-level fusion rather than identical classification tasks. A key limitation is that the modality indicators may encode dataset-availability patterns; future work will therefore evaluate ablations with and without `has_IR`, `has_V`, and `has_AUDIO`. Future studies will also consider sequence-aware validation, paired multimodal data, and external datasets to further assess generalization.

V. Conclusion

This work evaluated classical and *deep learning* models for multiclass aerial target classification using infrared, visible, and acoustic sensors under missing-modality conditions.

The best performance was achieved by IR+V+AUDIO early fusion with *HistGradientBoosting*, reaching BA of 0.9005, macro-F1 of 0.9002, and weighted-F1 of 0.9082. AUDIO was the strongest unimodal source, while classical models outperformed the evaluated deep architectures.

HistGradientBoosting provided the best performance–cost trade-off, whereas *Extra Trees* offered faster training with comparable performance. Future work will investigate modality-indicator ablations, attention-based fusion, sequence-aware validation, paired and external multimodal datasets, and improved analysis of the feature-extraction stage to strengthen reproducibility and generalization.

References

[1] B. Taha A. Shoufan, "Machine learning-based drone detection and classification: State-of-the-art in research," *IEEE Access*, vol. 7, pp. 138 669–138 682, 2019.

[2] Z. Liu, P. An, Y. Yang, S. Qiu, Q. Liu, X. Xu, "Vision-based drone detection in complex environments: A survey," *Drones*, vol. 8, no. 11, p. 643, 2024.

[3] F. Svanström, C. Englund, F. Alonso-Fernandez, "Real-time drone detection and tracking with visible, thermal and acoustic sensors," in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 7265–7272.

[4] F.-L. Chiper, A. Martian, C. Vladeanu, I. Marghescu, R. Craciunescu, O. Fratu, "Drone detection and defense systems: Survey and a software-defined radio-based solution," *Sensors*, vol. 22, no. 4, p. 1453, 2022.

[5] A. Famili, A. Stavrou, H. Wang, J.-M. J. Park, R. Gerdes, "Securing your airspace: Detection of drones trespassing protected areas," *Sensors*, vol. 24, no. 7, p. 2028, 2024.

[6] A. Coluccia, G. Parisi, A. Fascista, "Detection and classification of multirotor drones in radar sensor networks: A review," *Sensors*, vol. 20, no. 15, p. 4172, 2020.

[7] B. Taha A. Shoufan, "Machine learning-based drone detection and classification: State-of-the-art in research," *IEEE Access*, vol. 7, pp. 138 669–138 682, 2019.

[8] S. Samaras, E. Diamantidou, D. Ataloglou, N. Sakellariou, A. Vafeiadis, V. Magoulianitis, A. Lalas, A. Dimou, D. Zarpalas, K. Votis, P. Daras, D. Tzovaras, "Deep learning on multi sensor data for counter uav applications—a systematic review," *Sensors*, vol. 19, no. 22, p. 4837, 2019.

[9] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, J. Chen, "Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 68–74, 2018.

[10] C. Aker S. Kalkan, "Using deep networks for drone detection," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[11] M. Saqib, S. D. Khan, N. Sharma, M. Blumenstein, "A study on detecting drones using deep convolutional neural networks," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–5.

[12] E. Unlu, E. Zenou, N. Riviere, P.-E. Dupouy, "Deep learning-based strategies for the detection and tracking of drones using several cameras," *IPSJ Transactions on Computer Vision and Applications*, vol. 11, no. 7, pp. 1–13, 2019.

[13] S. Singha B. Aydin, "Automated drone detection using yolov4," *Drones*, vol. 5, no. 3, p. 95, 2021.

[14] A. Coluccia, A. Fascista, A. Schumann, L. Sommer, A. Dimou, D. Zarpalas, M. Mendez, D. De la Iglesia, I. Gonzalez, J.-P. Mercier *et al.*, "Drone vs. bird detection: Deep learning algorithms and results from a grand challenge," *Sensors*, vol. 21, no. 8, p. 2824, 2021.

[15] A. Coluccia, A. Fascista, L. W. Sommer, A. Schumann, A. Dimou, D. Zarpalas, "The drone-vs-bird detection grand challenge at icassp 2023: A review of methods and results," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 766–779, 2024.

TABLE V

BEST RESULT OBTAINED BY MODEL

Model	Family	Best scenario	Mode	BA	Macro-F1	Weighted-F1	Training (s)
HistGradientBoosting	Classical	IR+V+AUDIO	Early concat missing	**0.9005**	**0.9002**	**0.9082**	32.96
LogReg meta	Classical	IR+V+AUDIO	Late stacking	0.8978	0.8980	0.9067	180.28
Extra Trees	Classical	IR+V+AUDIO	Early concat missing	0.8912	0.8922	0.9008	1.32
Linear SVM	Classical	IR+V+AUDIO	Unimodal AUDIO	0.8888	0.8890	0.8872	0.06
Random Forest	Classical	IR+V+AUDIO	Early concat missing	0.8819	0.8803	0.8904	1.20
Logistic Regression L2	Classical	AUDIO	Unimodal AUDIO	0.8815	0.8800	0.8800	0.05
MLPClassifier	Classical	IR+V+AUDIO	Unimodal AUDIO	0.8339	0.8312	0.8329	0.07
ResNet-50 feature-map	Deep	IR+V+AUDIO	Deep early concat missing	0.8232	0.8236	0.8374	114.37
Feature Transformer	Deep	IR+V+AUDIO	Deep early concat missing	0.7984	0.7959	0.8091	22.27
Dummy prior	Classical	AUDIO	Unimodal AUDIO	0.3333	0.1667	0.1667	0.01

TABLE VI

COMPUTATIONAL EFFORT OF THE MAIN EVALUATED MODELS

Family	Model	Scenario / mode	Training (s)	Inference (ms/sample)	Epochs / best	Parameters	Size (MB)
Classical	HistGradientBoosting	IR+V+AUDIO / Early concat	32.96	3.95	–	–	7.63
Classical	LogReg meta	IR+V+AUDIO / Late stacking	180.28	–	–	–	–
Classical	Extra Trees	IR+V+AUDIO / Early concat	1.32	4.75	–	–	8.03
Classical	Linear SVM	IR+V+AUDIO / Unimodal AUDIO	0.06	0.56	–	–	0.06
Classical	Random Forest	IR+V+AUDIO / Early concat	1.20	4.09	–	–	3.22
Classical	Logistic Regression L2	AUDIO / Unimodal AUDIO	0.05	0.47	–	–	0.01
Classical	MLPClassifier	IR+V+AUDIO / Unimodal AUDIO	0.07	0.63	–	–	1.44
Deep	ResNet-50 feature-map	IR+V+AUDIO / Deep early concat	114.37	4.30	25.8/15.8	23.51 M	89.99
Deep	Feature Transformer	IR+V+AUDIO / Deep early concat	22.27	**0.29**	31.2/21.2	0.42 M	1.60
Classical	Dummy prior	AUDIO / Unimodal AUDIO	0.01	0.46	–	–	0.002

Note: decimal epoch values indicate averages over experimental repetitions.

TABLE VII

COMPARISON WITH SVANSTRÖM ET AL. [3]

Source	Scenario	Method	Metric	Value	Observation
Svanström et al.	IR	Thermal YOLOv2	F1-score	0.7601	Detection by distance
This work	IR	HistGradientBoosting	Macro-F1	0.8304	Feature-based classification
Svanström et al.	V	Visible YOLOv2	F1-score	0.7849	Detection by distance
This work	V	HistGradientBoosting	Macro-F1	0.7313	Feature-based classification
Svanström et al.	AUDIO	MFCC + LSTM	F1-score	**0.9323**	Three acoustic classes
This work	AUDIO	Extra Trees	Macro-F1	0.8810	Three acoustic classes
Svanström et al.	Fusion	Sensor outputs and confidences	Drone detection	0.7800	Detection opportunities
This work	IR+V+AUDIO	HistGradientBoosting	Macro-F1	**0.9002**	Five global classes

Note: the comparison is only indicative, since Svanström et al. [3] evaluate visual detection, acoustic classification, and decision-level fusion, whereas this work evaluates feature-based multiclass classification with missing-modality handling.

[16] A. Bernardini, F. Mangiatordi, E. Pallotti, L. Capodiferro, "Drone detection by acoustic signature identification," *Electronic Imaging*, vol. 2017, no. 10, pp. 60–64, 2017.

[17] S. Jeon, J.-W. Shin, Y.-J. Lee, W.-H. Kim, Y. Kwon, H.-Y. Yang, "Empirical study of drone sound detection in real-life environment with deep neural networks," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1858–1862.

[18] M. Z. Anwar, Z. Kaleem, A. Jamalipour, "Machine learning inspired sound-based amateur drone detection for public safety applications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2526–2534, 2019.

[19] S. Al-Emadi, A. Al-Ali, A. Al-Ali, "Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks," *Sensors*, vol. 21, no. 15, p. 4953, 2021.

[20] H. Liu, Z. Wei, Y. Chen, J. Pan, L. Lin, Y. Ren, "Drone detection based on an audio-assisted camera array," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017, pp. 402–406.

[21] S. Jamil, Fawad, M. Rahman, A. Ullah, S. Badnava, M. Forsat, S. S. Mirjavadi, "Malicious uav detection using integrated audio and visual features for public safety applications," *Sensors*, vol. 20, no. 14, p. 3923, 2020.

[22] E. Diamantidou, A. Lalas, K. Votis, D. Tzovaras, "Multimodal deep learning framework for enhanced accuracy of uav detection," in *Computer Vision Systems: 12th International Conference, ICVS 2019*, ser. Lecture Notes in Computer Science, vol. 11754. Springer, 2019, pp. 768–777.

[23] F. Svanström, C. Englund, F. Alonso-Fernandez, "Real-time drone detection and tracking with visible, thermal and acoustic sensors," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 7265–7272.

[24] F. Svanström, F. Alonso-Fernandez, C. Englund, "A dataset for multi-sensor drone detection," *Data in Brief*, vol. 39, p. 107521, 2021.

[25] S. Yuan, Y. Yang, T. H. Nguyen, T.-M. Nguyen, J. Yang, F. Liu, J. Li, H. Wang, L. Xie, "Mmaud: A comprehensive multi-modal anti-uav dataset for modern miniature drone threats," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 2745–2751.