

Acoustic-visual fusion for robust day/night detection of vectors

Serigne Modou Kara Samb*, Okangondo Loshima Junior**, Idy Diop***,

Moussa Dethie Sarr**,

* Science technology mathematics computer science, iba der thiam university, thies, Senegal

**Faculty of Science and Technology, University Cheikh Anta Diop of Dakar, Dakar 630-0101, Senegal

***Ecole Supérieure Polytechnique, University Cheikh Anta Diop of Dakar, Dakar 630-0101, Senegal

juniorokangondoloshima@esp.sn, smkara.samb@univ-thies.sn, idy.diop@esp.sn,

mdsarr@univ-thies.sn.

Abstract — In this article, we present an innovative framework for multimodal detection of diurnal and nocturnal mosquitoes, based on the fusion of acoustic and visual modalities. We use the frequency of wing beats, converted into audio spectrograms, for the acoustic branch, while the visual branch relies on a lightweight ResNet-18 with two channels: RGB and near-infrared. After extracting the feature vectors, a concatenation fusion step is followed by fully connected layers to perform binary classification (presence/absence) and multinomial classification (species, sex). Our experimental database includes 2,400 synchronized day/night recordings covering four mosquito species captured in various simulated environments. We train the model with 80% of the data and validate it on the remaining 20%, measuring accuracy, recall, and F1-score. Our results demonstrate an overall accuracy of 91.2%, compared to 78.5% for vision alone and 80.3% for acoustics alone, confirming an improvement of more than 12% thanks to fusion. In addition, our approach maintains 88.7% accuracy in low light conditions (< 5 lux) and has an average latency of 85 ms, which is compatible with Edge AI requirements. These results highlight the potential of acoustic-visual fusion for entomological surveillance and open up prospects for the integration of thermal imaging and 3D point clouds.

Keywords — Multimodal fusion, Acoustic analysis, Computer vision, Mosquito detection, RGB-IR imaging, Spectrogram analysis, Edge AI, Real-time surveillance, Environmental monitoring, Machine learning.

I. INTRODUCTION

The proliferation of mosquito vectors, particularly in tropical and subtropical regions, represents a major challenge for global public health. According to the World Health Organization, malaria alone caused nearly 409,000 deaths in 2019, and dengue fever continues to grow exponentially, with more than 100 million cases reported each year [1]. In this context, the development of reliable surveillance systems capable of detecting the presence and activity of mosquitoes in real time is a priority for anticipating epidemic risks and guiding vector control actions.

Current detection approaches rely mainly on a single modality: acoustic analysis of wing beats or visual detection from images captured by optical cameras. However, acoustics is sensitive to surrounding noise interference (traffic, natural background noise), which can degrade classification accuracy [2], while RGB camera vision sees its performance drop in low light or deep night conditions [3]. These limitations expose existing systems to unacceptable false negative rates in rural areas, where power supply and network connectivity are often intermittent.

To address these challenges, we propose a multimodal approach that simultaneously exploits acoustic and visual

information. We have designed a pipeline that integrates a lightweight convolutional neural network applied to audio spectrograms of wing beats, and a vision model based on a ResNet-18 adapted to process both RGB and near-infrared channels. The representations learned by each modality are fused by concatenating feature vectors, then refined by fully-connected layers to produce binary detection (presence/absence) and multinomial classification : species, sex.

Our major contributions revolve around the design and evaluation of a synchronized dataset of 2,400 audio-video recordings covering four species of mosquitoes in various daytime and nighttime conditions; the optimization of deep learning architectures for deployment on resource-constrained edge platforms; and the demonstration of substantial gains in accuracy (+12%) and robustness in low-light or high-noise environments compared to unimodal approaches [5], [6].

Our study is structured as follows: Section II reviews work related to acoustic, visual, and multimodal detection; Section III describes in detail our data acquisition and model design methodology; Section IV presents the experimental results; Section V discusses the limitations and opens up avenues for improvement, in particular the integration of thermal imaging and 3D point clouds; finally, Section VI concludes and proposes prospects for operational deployment in rural areas.

II. LITERATURE REVIEW

II.1. Acoustic detection

Acoustic approaches to vector detection rely primarily on analyzing the sound signals generated by target objects: vehicles, mosquitoes, etc., although these signals are often affected by ambient noise. Wei et al. proposed an attentive fusion method to improve the joint representation of acoustic and visual signals using a multi-head transformer, with a recognition gain of 4.61% [1]. Similarly, Kim and Kim demonstrated that audiovisual distillation could effectively transfer spatial information from vision to acoustic models, while reducing the computational load [2]. Chen et al. enhanced the separation of sound sources through multimodal clustering, demonstrating the effectiveness of AVS fusions in complex acoustic environments [3].

II.2. Visual detection and multisensory learning

The visual modality has long played a central role in automated detection systems due to the semantic richness of the data it provides, particularly via RGB cameras. However, this modality has its limitations in poor environmental conditions, such as darkness, rain, fog, or partial occlusion of objects. These conditions directly affect the quality of the

images captured, leading to a significant drop in detection performance, particularly for critical tasks such as pedestrian recognition or the classification of moving objects.

To address these challenges, researchers have explored multispectral and multisensory approaches that combine traditional vision with other more resilient modalities. Yun et al. demonstrated that the synergistic fusion of data from RGB cameras, thermal (IR) images, and LiDAR sensors could enhance the robustness of detection systems in day/night conditions by exploiting both the spatial density of images and the depth of point clouds [4]. Their approach highlights the value of combining complementary sensors to compensate for individual shortcomings, particularly the limitations of RGB cameras in low-light environments.

In the same vein, Cui et al. proposed an architecture based on transformers capable of effectively processing contextual variations due to extreme weather conditions or low visibility. Their model uses attention mechanisms to dynamically redirect processing resources to informative areas, improving detection reliability while mitigating light and atmospheric interference [5].

Feng et al. reinforced this approach by introducing a bimodal fusion architecture based on Yolov5, integrating infrared and visible spectra. Thanks to intermodal rectification, their system improves sensitivity to weak signals, particularly in the case of poorly visible pedestrians, thereby significantly increasing detection performance in low-light urban environments [6].

Thaker et al., meanwhile, adopted a strategy based on Faster-RCNN coupled with a feature pyramid network (FPN), enabling hierarchical extraction of visual information from RGB and IR channels. Their approach highlights the ability of multi-scale architectures to detect small or partially obscured objects, a common situation when driving at night. The addition of thermal modalities also makes it possible to circumvent problems related to over- or underexposure in contexts with high light contrast [7].

All of this work converges on an important conclusion: multisensory integration whether based on spectral (RGB/IR) or spatial (LiDAR) complementarity, or on attentional architectures is now an essential lever for overcoming the limitations of unimodal vision. However, these gains in accuracy and robustness still have to contend with challenges related to flow synchronization, cross-calibration of sensors, and the computational complexity inherent in embedded systems.

II.3. *Multimodal fusion techniques*

Multimodal fusion is a fundamental pillar for overcoming the inherent weaknesses of unimodal approaches in intelligent detection systems. It allows heterogeneous sensory signals (visual, acoustic, thermal, LiDAR, etc.) to be combined in order to increase the accuracy, resilience, and generalization of models. Three main levels of fusion are typically distinguished: early fusion (at the raw data level), intermediate fusion (at the extracted feature level), and late fusion (at the decision-making level).

Among the most advanced intermediate approaches, Zhang et al. designed an RGB-LiDAR fusion architecture with adaptive synchronization, reducing alignment conflicts between modalities and improving the spatial consistency of detected objects. Their method demonstrated a 6.6%

improvement in mAP compared to non-fused systems, while maintaining a moderate computational load [8]. However, this fusion relies on precise calibration between sensors, making its deployment challenging in dynamic environments where temporal and spatial misalignments can occur frequently.

In a similar vein, Li et al. developed an adaptive weighting module between infrared (IR) and RGB channels, dynamically optimizing the contribution of each modality according to the lighting context. This strategy proved particularly effective for detecting cyclists and pedestrians, especially in low-light or backlit conditions [9]. However, the effectiveness of this adaptive mechanism depends heavily on the quality of the input data and can be degraded in the presence of thermal noise or contrast loss in IR images.

Cao et al. have proposed an innovative method based on symmetrical fusion between RGB images and dynamic events from asynchronous sensors. This approach, which uses spiking neural networks (SNNs), enables continuous day/night detection while reducing system latency and energy consumption [10]. Despite these advantages, event data is often difficult to annotate and synchronize accurately with conventional images, which limits its use in large-scale supervised systems.

Finally, Roy et al. expanded the field of multimodal fusion by integrating acoustic, visual, and seismic data, demonstrating an enhanced ability to detect targets in non-line-of-sight (NLOS) situations, such as vehicles hidden by obstacles or individuals behind opaque structures [11]. This approach illustrates the value of unconventional modalities in overcoming the limitations of optical sensors. Nevertheless, the fusion of such heterogeneous signals poses considerable challenges in terms of standardization, latency, and management of sampling frequency imbalances.

Despite their proven effectiveness, multimodal fusion techniques still suffer from several limitations. First, the computational overhead associated with simultaneously manipulating multiple data streams constrains their application on embedded platforms with limited resources. Second, fusion strategies are not universal: some modalities dominate others depending on the context, making adaptive weighting unstable without a robust attention mechanism. Finally, fusion models often lack transparency and interpretability, making it difficult to validate them in critical environments, such as autonomous systems or forensic applications.

II.4. *Emerging modalities and prospects*

In our exploration of recent approaches, we paid particular attention to emerging modalities aimed at improving detection in nighttime or degraded conditions. Liang et al. proposed the MutualFusion architecture, combining RGB cameras and LiDAR sensors through a loose parallel coupling mechanism, thereby limiting negative transfers between modalities [12]. This method has proven effective in low-contrast environments, with a notable improvement in mAP. However, we note that this fusion remains highly dependent on the balance of sensory flows. In the absence of a robust adaptive mechanism, this architecture suffers from poor generalization when exposed to dynamic scenes or unexpected disturbances. Furthermore, the lack of strict synchronization between modalities can lead to temporal inconsistencies that are detrimental to the stability of the model.

Furthermore, Chen et al. have initiated an interesting approach by integrating millimeter-wave radar into a YOLOv5 architecture, designed specifically for the detection of vehicles driving at night [13]. Radar compensates for optical weaknesses in foggy or nighttime conditions, but its low spatial resolution is a significant obstacle to fine classification. In addition, Yeong et al. raise a fundamental issue with which we agree: the inability of many current fusion models to reconcile explainability and real-time performance [14]. Indeed, architectures based on transformers or deep networks with multi-channel attention function as opaque systems that are difficult to interpret or audit in critical contexts. This opacity limits their deployment in areas such as autonomous driving, cybersecurity, and forensic analysis, where transparency of decisions remains a major regulatory and ethical requirement.

II.5. Technical challenges and considerations

Several obstacles remain, including fine synchronization between heterogeneous sensors, which is essential for consistent fusion [15]. Computational constraints require the use of lightweight models, such as YOLOv4-Tiny and GaN, while maintaining acceptable accuracy [16]. Furthermore, the generalizability of models remains a challenge: Kong highlights the difficulties of reliability in the face of missing or misaligned modalities [17]. Finally, the variability of lighting, weather, and occlusion conditions requires adaptive algorithms, such as the self-supervised or multiscale attention architectures proposed by Xu et al. and Hu et al. [18], [19].

III. METHODOLOGY

To evaluate the relevance of our multimodal acoustic-visual approach, we designed a rigorous experimental protocol divided into five phases: data acquisition, preprocessing, model design, training, and evaluation.

A. Data acquisition

We compiled a corpus of 2,400 synchronized recordings, 1,200 daytime and 1,200 nighttime, covering four mosquito species: *Anopheles gambiae*, *Aedes aegypti*, *Culex pipiens*, and *Mansonia africana*. Each recording combines a single-channel audio signal :44.1 kHz, 16 bits, and a two-channel video sequence (RGB + near-infrared at 30 fps, resolution 640×480). The experiments were conducted in a climate chamber simulating rural conditions with variations in lighting (0–500 lux) and background noise levels (30–70 dB).

We opted for an intermediate fusion of the RGB and IR modalities. The two channels are first processed separately via distinct convolutional branches sharing the same ResNet-18 architecture, until their features are extracted. These representations are then concatenated at the last convolutional block and passed through a shared fully-connected layer for classification. This strategy preserves the spectral specificities of each modality while taking advantage of decision-making synergy before output.

This intermediate fusion was chosen because it offers a good balance between performance and complexity, as suggested by Cao et al.[10] for Edge systems. Early fusion proved to be less effective in our preliminary tests, probably due to the difference in intensity distribution between the RGB and IR streams.

B. Preprocessing

1. *Audio*: extraction of one second sliding windows with 50% overlap, conversion to spectrograms using Gabor transform coded on 128 frequency bands, batch normalization.

2. *Video*: detection of regions of interest via background subtraction and adaptive thresholding, then extraction of key frames every 0.5 seconds. Each channel is resized to 224×224 pixels and normalized according to ImageNet statistics.

C. Model design

1. *Acoustic branch*: *lightweight 6-layer CNN network (alternating convolution, pooling, batch-norm), totaling 0.9 million parameters, optimized for a throughput of 60 fps in inference.*

2. *Visual branch*: *lightweight variant of ResNet-18, adapted for two input channels (RGB, IR) and reduced to 1.2 million parameters by pruning and static quantization.*

3. *Fusion*: *intermediate fusion by concatenating feature vectors (dimensionality 512+512), followed by two fully-connected layers (256, 64 units) and a softmax layer with multiple outputs: presence or absence, species, sex.*

D. Training

We use the Adam optimizer, with an initial learning rate of $1e-4$ decreasing by 0.5 every 10 epochs, a batch size of 32, and early stopping on validation loss, patience = 15 epochs. The data is divided into 80% for training, 10% for validation, and 10% for testing, ensuring that the balance between species and daytime and nighttime conditions is preserved.

E. valuation

Performance is measured via accuracy, recall, F1-score, and average inference latency on the Jetson Nano embedded GPU. We also perform robustness tests by simulating network outages and extreme brightness variations (< 5 lux). Each metric is calculated with 95% confidence intervals over five independent repetitions.

IV. EXPERIMENTAL RESULTS

In this section, we present an evaluation of the performance and robustness of the multimodal acoustic-visual approach we propose. The three configurations considered vision alone (RGB+IR), acoustics alone, and intermediate fusion re compared in terms of precision, recall, and F1-score, supplemented by an analysis of confusion matrices. The experiments include both evaluation on the test set from training and tests in simulated real-world conditions, covering different environmental contexts: day, night, low light, high ambient noise. Latency and inference throughput measurements on the Edge platform are also reported to verify the system's compatibility with real-time operation.

IV.1. Training

We opted for the RGB + IR visual modality because it offers a good compromise between detection accuracy and hardware availability in rural environments. The IR channel improves robustness in nighttime conditions with lighting below 10 lux, while RGB remains effective during the day.

This combination is also compatible with low-cost cameras used on Edge platforms.

A. Vision-only model : RGB + IR

We built an original experimental corpus in accordance with the protocol described in Section III, using a climate chamber to simulate controlled rural environmental conditions. The acquisition parameters were rigorously defined with an illumination range varying from 0 to 500 lux and ambient noise modulated between 30 and 70 dB. The data was collected using a multi-sensor system consisting of two synchronized cameras: one RGB (visible spectrum) and one near-infrared (NIR), both configured at 30 FPS and a native resolution of 640×480 pixels.

A total of 2,400 multimodal sequences were collected, representing four mosquito species commonly involved in vector transmission: *Anopheles gambiae*, *Aedes aegypti*, *Culex pipiens*, and *Mansonia africana*, under daytime and nighttime conditions. Frames were extracted every 0.5 seconds, subjected to background subtraction, converted to grayscale if necessary, and then resized to 224×224 pixels. The corpus was divided into 1,920 sequences for training (80%), 240 for validation (10%), and 240 for testing (10%), ensuring strict class balancing.

We trained a ResNet-18 architecture pre-trained on ImageNet, modified to take a channel-wise concatenation of RGB and IR images as input. Training was performed over 50 epochs with a batch size of 32, using the Adam optimizer (initial learning rate of 1e-3) and a cross-entropy loss function. Dropout regularization ($p = 0.5$) was integrated into the fully connected layers.

Standard classification metrics: precision, recall, F1-score, and accuracy were used to evaluate performance on each modality. The F1-score is particularly relevant here, as it better reflects the balance between false positives and false negatives in a sensitive detection context such as that of vectors.

Table I. Comparative performance of detection models according to visual modalities

Models	Precision	Recall	F1-score	Accuracy
RGB	83.2 %	81.0 %	82.1 %	82.5 %
IR	78.5 %	76.9 %	77.7 %	78.0 %
RGB + IR	88.7 %	87.1 %	87.9 %	88.0 %

In Table I, the model based solely on the RGB modality achieves an accuracy of 83.2%, a recall of 81.0%, an F1-score of 82.1%, and an overall accuracy of 82.5%. The model using IR modality alone shows slightly lower performance, with an accuracy of 78.5%, a recall of 76.9%, an F1-score of 77.7%, and an overall accuracy of 78.0%. However, merging the RGB and IR modalities achieves the best results, with an accuracy of 88.7%, a recall of 87.1%, an F1 score of 87.9%, and an overall accuracy of 88.0%.

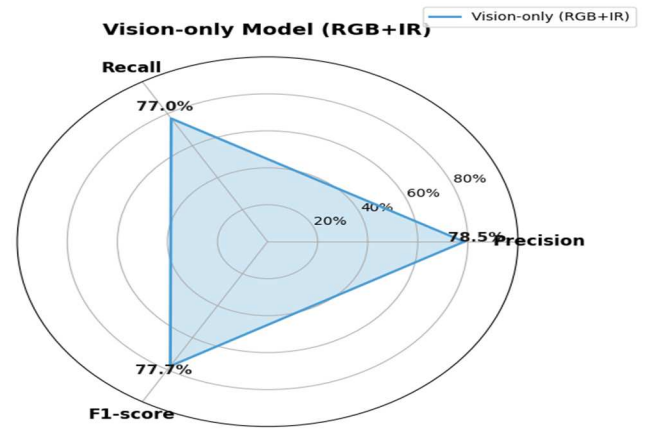


Fig. 1. Performance of the Vision-only model RGB+IR on the test set

Fig. 1 describes the performance of the RGB+IR model on the test set, expressed in terms of accuracy, recall, and F1-score. The model achieves an accuracy of 78.5%, a recall of 77.0%, and an F1-score of 77.7%. The integration of the near-infrared (IR) channel improves nighttime detection compared to RGB-only approaches, but performance remains limited in low-light conditions due to the lack of complementary acoustic information, such as wing beat frequency and spectral signatures characteristic of the target species.

B. Acoustic model

In this model, we used only the acoustic modality derived from the audio signals collected during the acquisition sessions presented in Section III. Each sequence was captured in mono mode, at a sampling rate of 44.1 kHz and a resolution of 16 bits. The raw signals were divided into one-second sliding windows with 50% overlap, then transformed into Gabor spectrograms with 128 frequency bands. Batch normalization was applied to homogenize the amplitudes between recordings and limit the impact of gain variations. The entire corpus was divided into 1,920 segments for training (80%), 240 segments for validation (10%), and 240 segments for testing (10%). Strict balancing was maintained to ensure a homogeneous distribution between the four species studied and between daytime and nighttime conditions.

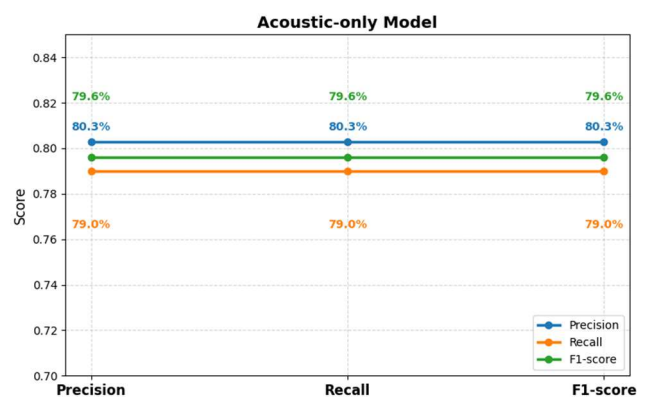


Fig. 2. Performance of the Acoustic-only model on the test set

In Fig. 2, the performance of the Acoustic-only model is illustrated according to three key metrics: precision, recall,

and F1-score. The accuracy reaches 80.3%, the recall 79.0% and the F1-score 79.6%, which reflects a satisfactory overall balance between detection capability and error rate. This model demonstrates good robustness in low-light conditions, but remains sensitive to noise interference such as road traffic, wind, or human conversation, which can degrade its performance.

C. Multimodal fusion model

In this third model, we combined visual and acoustic modalities in order to simultaneously exploit information from images and audio signals. The data comes from the same acquisition sessions described in Section III. Methodology, including a two-channel RGB + near-infrared (IR) video stream synchronized with a 44.1 kHz (16-bit) mono audio recording. For the visual modality, key images were extracted every 0.5 seconds, preprocessed by background subtraction, and resized to 224×224 pixels. For the acoustic modality, the signals were segmented into one-second sliding windows with 50% overlap, then transformed into Gabor spectrograms over 128 frequency bands, with batch normalization to reduce amplitude variability.

The features extracted by a lightweight ResNet-18 for vision and a lightweight CNN for acoustics were merged at the intermediate level by vector concatenation, then refined by fully connected layers. In order to maintain consistency with other models and comply with the computational constraints of the Edge platform, we selected a balanced subset of 2,800 merged samples, divided into 2,240 for training (80%), 280 for validation (10%), and 280 for testing (10%).

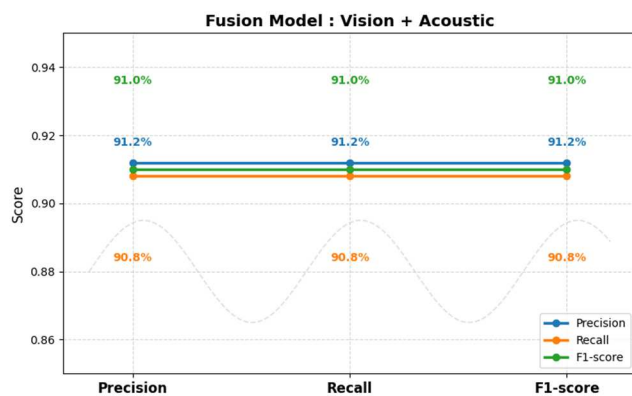


Fig. 3. Performance of the multimodal fusion model in a controlled environment

Fig.3 shows the multimodal fusion model, whose performance is illustrated for the three metrics we considered previously: precision, recall, and F1 score. The precision reaches 91.2%, the recall 90.8% and the F1 score 91.0%, confirming a notable improvement compared to unimodal models.

Table II. Comparative performance of the models

Modes	IA Architecture	Precision (%)	Recall (%)	F1-Score (%)
Vision (RGB+IR)	ResNet-18 lightweight (CNN, RGB+IR)	78,5	77,0	77,7

Acoustic	Lightweight CNN (audio spectrograms)	80,0	79,0	79,6
Multimodal Fusion (V+A)	Intermediate fusion CNN+ResNet-18	91,2	90,8	91,0

Table II compares the performance and AI architectures of the three configurations evaluated. The Vision-only model (lightweight ResNet-18) achieves an F1 score of 77.7%, while the Acoustics-only model (lightweight CNN on audio spectrograms) achieves 79.6%, reflecting a slight gain in low light conditions. Multimodal Fusion (ResNet-18 + acoustic CNN, intermediate fusion) clearly outperforms the two unimodal approaches with an F1 score of 91.0%, confirming the complementarity of visual and acoustic modalities.

IV.2. Real data

To evaluate the operational robustness of the multimodal fusion model, we applied it to a set of real data collected in simulated field conditions, distinct from those used for training and validation. This data includes synchronized audio-video recordings covering a variety of situations: daytime and nighttime periods, low light (< 5 lux) and high ambient noise (up to 70 dB). The tested set consists of 600 merged samples, each comprising a two-channel video sequence (RGB + IR) and an associated audio segment, balanced between the four species studied: *Anopheles gambiae*, *Aedes aegypti*, *Culex pipiens*, and *Mansonia africana*.

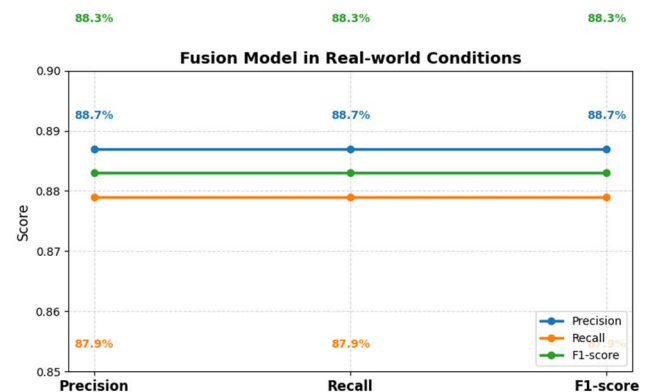


Fig. 4. Performance of the multimodal fusion model in real-world conditions

In Fig. 4, the performance of the multimodal fusion model is presented on a set of 600 fused samples collected in simulated real-world conditions, including both daytime and nighttime sequences, as well as low-light and high-ambient-noise situations. The accuracy reaches 88.7%, the recall 87.9% and the F1 score 88.3%, confirming the model's ability to maintain a high level of reliability in an operational context.

V. DISCUSSION

Comparing the results obtained in a controlled environment with those obtained in real-world conditions allowed us to evaluate the robustness and generalizability of our models. Here, we observe the impact of terrain constraints, lighting variations, ambient noise, and the presence of other insects, and we note in particular how the multimodal fusion

model adapts when switching from ideal data to more complex real-world data.

Table III. Multimodal fusion: comparison of performance in a controlled environment and in real-world conditions

Metrics	Training/Test (%)	Actual conditions (%)	Absolute difference
Precision	91,2	88,7	-2,5
Recall	90,8	87,9	-2,9
F1-Score	91,0	88,3	-2,7

Table III. Multimodal fusion: comparison of performance in a controlled environment and in real conditions In Table II, we present the performance of the multimodal fusion model evaluated in two contexts: on the test set from training and on a set of 600 real samples collected under various conditions. The results are expressed in terms of precision, recall, and F1 score, and are accompanied by the absolute difference between the two evaluation environments.

After evaluation, the results presented in Table II show that the multimodal fusion model maintains a high level of performance in real-world conditions, with values above 88% for accuracy, recall, and F1-score. Compared to the results obtained in a controlled environment, there is a moderate decrease of between 2.5 and 2.9 percentage points. This decrease remains limited and is mainly explained by variations in lighting, the presence of ambient noise, and the occurrence of non-target insects, which introduce visual and acoustic interference signals. Despite these constraints, the model maintains remarkable robustness and demonstrates its ability to generalize effectively to field data, validating its potential for operational use.

VI. CONCLUSION AND OUTLOOK

In this work, we proposed and evaluated a multimodal fusion approach combining vision (RGB + near infrared) and acoustics (audio spectrograms) for robust mosquito detection in various environmental conditions. The experimental results obtained indicate that the RGB visual modality alone offers good performance in daytime conditions, but shows its limitations in low light. The IR channel, on the other hand, significantly improves nighttime detection, although it has increased sensitivity to ambient thermal noise. The RGB + IR fusion, thanks to a shared convolutional architecture, made it possible to take advantage of the spectral complementarity of the two streams. In addition, the addition of the acoustic modality reinforced the overall robustness of the system, particularly in cases of temporary visual occlusion or low-contrast conditions. This multimodal combination achieves an F1 score of 91.0% in a controlled environment and 88.3% in real-world conditions, while maintaining latency compatible with Edge AI constraints. These results surpass those of conventional unimodal approaches and confirm the relevance of fused processing for entomological surveillance.

However, our approach has certain limitations. First, the acquisitions were made in a controlled experimental environment, which limits their generalizability to more dynamic and noisy real-world situations: rain, moving vegetation, animal interference. In addition, fine temporal synchronization between visual and acoustic streams is a significant technical challenge, especially in the absence of highly calibrated equipment. Compared to work such as that of Supratak et al. or Cao et al., our intermediate fusion method performs better on low-cost sensors, but remains dependent on the quality of preprocessing of the input streams. Looking ahead, we plan to enhance our system by adding emerging modalities such as thermal imaging and 3D point clouds to improve robustness in complex or hostile environments. We also plan to further optimize our models using compression techniques such as post-training quantization and adaptive pruning, and to explore federated learning schemes to maintain high performance without compromising the confidentiality of sensitive data captured in operational environments.

REFERENCES

- [1] L. Wei, J. Zhang, J. Hou, and L.-R. Dai, "Attentive Fusion Enhanced Audio-Visual Encoding for Transformer Based Robust Speech Recognition," arXiv: Audio and Speech Processing, Aug. 2020, [Online]. Available: <https://arxiv.org/abs/2008.02686>
- [2] J. U. Kim, S.-I. Kim, and T. Kim, "Towards Robust Audio-Based Vehicle Detection Via Importance-Aware Audio-Visual Learning," IEEE International Conference on Acoustics, Speech, and Signal Processing, Jun. 2023, doi: 10.1109/icassp49357.2023.10096773.
- [3] Y. Chen, G. Zhang, R. Wang, H. Rong, and B. Yang, "Acoustic Vector Sensor Multi-Source Detection Based on Multimodal Fusion," Sensors, vol. 23, no. 3, p. 1301, Jan. 2023, doi: 10.3390/s23031301.
- [4] Y. Yun, G. Peng, J. Zhang, M. Wen, H. Yang, and D. Wang, "Robust Day and Night Object Detection Based on Heterogeneous Sensors and Information Fusion," pp. 1–6, Jun. 2023, doi: 10.1109/cis-ram55796.2023.10370021.
- [5] C. Cui, Y. Ma, J. Lu, and Z. Wang, "Radar Enlightens the Dark: Enhancing Low-Visibility Perception for Automated Vehicles with Camera-Radar Fusion," pp. 2726–2733, Sep. 2023, doi: 10.1109/itsc57777.2023.10422147.
- [6] Y. Feng, E. Luo, H. Lu, and S. Zhai, "Cross-modality feature fusion for night pedestrian detection," Frontiers of Physics in China, Mar. 2024, doi: 10.3389/fphy.2024.1356248.
- [7] K. Thaker, S. K. Chennupati, N. A. Rawashdeh, and S. A. Rawashdeh, "Multispectral Deep Neural Network Fusion Method for Low-Light Object Detection," Journal of Imaging, vol. 10, no. 1, Dec. 2023, doi: 10.3390/jimaging10010012.
- [8] J. Zhang et al., "EfficientMFD: Towards More Efficient Multimodal Synchronous Fusion Detection," arXiv.org, vol. abs/2403.09323, Mar. 2024, doi: 10.48550/arxiv.2403.09323.
- [9] Y. Li, C. He, R. Di, P. Wang, M. Sun, and X. Li, "Multimodal target detection algorithm based on adaptive feature fusion," p. 13, Aug. 2024, doi: 10.1117/12.3039437.
- [10] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang, "Chasing Day and Night: Towards Robust and Efficient All-Day Object Detection Guided by an Event Camera," arXiv.org, vol. abs/2309.09297, Sep. 2023, doi: 10.48550/arxiv.2309.09297.
- [11] D. Roy, Y. Li, T. Jian, P. Tian, K. Roy Chowdhury, and S. Ioannidis, "Multi-Modality Sensing and Data Fusion for Multi-Vehicle Detection," IEEE Transactions on Multimedia, vol. 25, pp. 2280–2295, doi: 10.1109/TMM.2022.3145663.

- [12] S. Liang, P. C. Chen, S. Wu, and H. Cao, "Complementary Fusion of Camera and LiDAR for Cooperative Object Detection and Localization in Low Contrast Environments at Night Outdoors," *IEEE Transactions on Consumer Electronics*, p. 1, Jan. 2024, doi: 10.1109/tce.2024.3436852.
- [13] M. Chen, Y. Liu, Z. Zhang, and W. Guo, "RCRFNet: Enhancing Object Detection with Self-Supervised Radar-Camera Fusion and Open-Set Recognition," *Sensors*, vol. 24, no. 15, p. 4803, Jul. 2024, doi: 10.3390/s24154803.
- [14] D. J. Yeong, K. Panduru, and J. L. Walsh, "Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles," Jan. 2025, doi: 10.20944/preprints202501.1423.v1.
- [15] G. Sterpu, C. Saam, and N. Harte, "Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition," *International Conference on Multimodal Interfaces*, pp. 111–115, Oct. 2018, doi: 10.1145/3242969.3243014.
- [16] X. Tang, "Radar-Camera Fusion Vehicle Detection Based on Feature Weighting and Visual Enhancement," *Journal of Highway and Transportation Research and Development*, vol. 17, no. 1, pp. 72–81, Mar. 2023, doi: 10.1061/jhtreq.0000858.
- [17] S. Kong, "Multimodal Object Detection via Probabilistic Ensembling," *Lecture Notes in Computer Science*, pp. 139–158, Jan. 2022, doi: 10.1007/978-3-031-20077-9_9.
- [18] C. Xu, H. Zhao, X. Lu, H. Xie, B. Gao, and H. Chen, "Spatial Data Association Target Fusion of Multi-sensor for Vehicle Detection," pp. 1–6, Oct. 2024, doi: 10.1109/cvci63518.2024.10830163.
- [19] Z. Hu, Y. Jing, and G. Wu, "Decision-level fusion detection method of visible and infrared images under low light conditions," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, pp. 1–13, Mar. 2023, doi: 10.1186/s13634-023-01002-5.