AI for Rural Fires Detection

Magda Costa - GECAD/ISEP

Advisor: Eva Maia
Co-Advisor: Ivone Amorim

Youtube Video
Test Images on the developed Models

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
3 de Julho de 2025

I. Abstract

This project explored the use of artificial intelligence based models for detecting rural fires using images captured by Unmanned Aerial Vehicles (UAVs). Its primary goal was to develop and assess models capable of generalizing to unseen datasets, which is a key requirement for real-world deployment. The work reproduced and tested two state-of-the-art convolutional neural networks using several publicly available UAV datasets. To improve generalization, the methodology included baseline reproduction, cross-dataset evaluation, data augmentation, and dataset combination. Results highlighted key factors influencing generalization, such as dataset diversity, and revealed the strengths and limitations of each model. The findings support the development of more robust and adaptable wildfire detection systems.

II. Introduction

Rural fires have emerged as a rapidly expanding global concern, especially with the onset of climate change, intensifying the frequency and intensity of such events [1] [2]. These fires result in devastating ecological and economic impairment and pose serious threats to human life and property [2] [3]. Early fire detection plays a crucial role in minimizing such impacts, however, traditional monitoring systems, such as human patrols, watchtowers, and satellite observation, often struggle to accurately and promptly identify fire outbreaks due to limitations in spatial coverage and delayed response times [4]. As such, there is a growing need for more responsive, accurate, and scalable fire detection systems that can operate effectively in real-world environments.

Recent advances in Artificial Intelligence (AI) and computer vision offer promising solutions to address these challenges [3]. In particular, Unmanned Aerial Vehicles (UAVs) equipped with intelligent vision systems provide a flexible and efficient means of monitoring vast and remote landscapes [2].

Using images captured by UAVs, this study examined the application of convolutional neural networks (CNNs) for detecting rural fires. In particular, two cutting-edge CNN architectures were chosen from recent research, faithfully replicated, and subjected to a series of experiments to evaluate their performance and generalization ability. Several publicly available datasets were used to train and evaluate these models, with a focus on assessing their robustness in cross-dataset situations. To enhance generalization, the project also explored the effects of data augmentation and dataset combination strategies. These steps were designed to emulate real-world variability and test whether exposure to heterogeneous data could improve model adaptability. The main contributions of this project are: a thorough evaluation and replication of two cutting-edge CNN models for UAV-based wildfire detection, a comprehensive assessment of their generalization abilities across several datasets, an analysis of the effects of data augmentation and dataset merging on model robustness, and the publication of all developed code and trained models in an open-access GitHub repository [5] to promote transparency, reproducibility, and further research in the field.

This document is structured as follows: Section III provides a review of the state of the art in fire detection systems, with an emphasis on AI-based approaches and the integration of UAVs. Section IV details the datasets used and their selection criteria. Section V outlines the methodology adopted, including model implementation, training procedures, and evaluation strategies. Section VI presents and discusses the experimental results. Finally, Section VII concludes with a summary of the findings and potential directions for future work.

III. State of the Art

A thorough review of the state of the art is essential to contextualize the current work within existing research and to identify the main technological and scientific advances in wildfire detection. This section outlines the progression from traditional fire monitoring methods to more recent approaches that incorporate AI and aerial systems, including a review of public datasets, and the main existing limitations.

A. Overview of Forest Fire Detection Technologies

Wildfires represent a chronic environmental and socioeconomic threat exacerbated by climate and land-use changes in recent decades [2] [3]. Wildfire detection systems traditionally rely on surveillance towers, human patrols, and satellite-based remote sensing. These methods, although fundamental, are prone to severe limitations such as limited spatial resolution, time lag in detection, and ineffectiveness under poor visibility conditions (e.g., smoke, fog, or cloud cover) [2] [6].

Satellite-based systems, particularly those utilizing instruments such as MODIS or VIIRS, provide broad coverage but are hampered by low revisit frequencies and are also vulnerable to atmospheric interferences [7]. Thermal sensors and ground-based cameras have also been adopted as forest surveillance tools and are often part of centralized monitoring networks. These systems tend to be static, which leads them to have constrained

fields of view and to require substantial infrastructure investments [2] [3].

Recent advances have introduced Internet of Things (IoT) technology, integrating wireless communications and sensor networks to observe environmental factors like temperature, humidity, and gas concentrations. Although these systems offer real-time local data, they are restricted to large-scale applications, i.e., maintenance at remote locations and susceptibility to false alarms due to natural variability [4].

To overcome these challenges, UAVs have emerged as a dynamic and cost-effective tool in wildfire detection [4] [3]. UAVs can capture high-resolution images, navigate challenging terrains, and provide immediate feedback in active fire zones. Combined with advanced AI techniques, they offer flexibility, which increases early detection by leaps and bounds and enables preventive fire responses and management programs [4] [3] [6]. However, despite these advantages, UAVS-bases systems still face limitations. These include restricted flight duration due to limited battery capacity, and constrained onboard processing capabilities that hinder real-time analysis [3].

B. AI Techniques for Fire Detection

The increasing frequency and intensity of wildfires have generated significant interest in using AI for fire detection and management [3]. AI techniques, particularly Machine Learning (ML) and Deep Learning (DL) have been widely explored for their ability to analyze complex patterns in visual and environmental data, enabling early fire detection and predicting fire behavior [6] [7].

CNNs have been among the most widely adopted architectures for image-based fire detection due to their ability to extract hierarchical spatial features. For instance, Mowla et al. (2024) [6] proposed an Adaptive Hierarchical Multi-Headed CNN with a Modified Convolutional Block Attention Module (AHMHCNN-mCBAM), which integrates attention mechanisms and temporal mechanisms (GRU and BiLSTM) to improve the accuracy of fire detection in aerial images. Pursuing a complementary objective, Ramadan et al. [4] developed a lightweight model suitable for real-time inference on edge devices. The study implements a convolutional neural network architecture optimized through Neural Architecture Search. The resultant CNN features a simplified design that reduces the number of parameters and depth, striking a purposeful balance between high classification performance and computational efficiency.

Metaheuristic optimization has also been employed to enhance the performance of deep networks. Rajalakshmi et al. (2023) [8] introduced a hybrid model combining Capsule Networks with Biogeography-Based Optimization and Deep Neural Networks, yielding improved classification results using UAV imagery, exceeding an accuracy of 99%.

Transformer-based architectures are also gaining ground. Chaturvedi et al. (2024) [7] developed an ultra-lightweight convolution-transformer hybrid model for smoke detection, achieving over 99% accuracy on multiple datasets while remaining computationally efficient, which is a critical feature for deployment on resource-constrained UAVs.

Surveys such as those by Giannakidou et al. (2024) [2] and Boroujeni et al. (2024) [3] highlight the increasing use of Reinforcement Learning (RL) for decision-making tasks in active-fire scenarios and the incorporation of AI across the wildfire lifecycle, addressing topics from prevention and detection to post-fire restoration.

Despite these advances, challenges remain. Many AI models lack generalization capabilities across varying environmental conditions and datasets [6]. Moreover, high false positive rates and computational demands limit the real-time applicability of some deep models in field conditions [9].

C. Use of UAVs in Wildfire Management

UAVs have emerged as essential tools in the domain of wildfire management, offering significant advantages in data acquisition, flexibility, and cost-effectiveness [4]. Their capacity to navigate hazardous or remote areas and to collect real-time, high-resolution imagery has transformed fire detection, monitoring, and response strategies [3] [6].

Boroujeni et al. (2024) [3] provided a comprehensive review of UAV integration across the wildfire lifecycle, identifying key contributions in three phases: pre-fire (fuel and risk assessment), active fire (detection, monitoring, and suppression), and post-fire (damage evaluation and recovery planning). The study highlights how AI-enabled UAVs, capable of autonomous operation and real-time data processing, can significantly enhance situational awareness and decision-making during fire events.

Ramadan et al. (2024) [4] described a UAV-IoT system featuring a novel CNN-based fire detection model embedded in drones, capable of identifying fires within 1–5 minutes after ignition and accurately geolocating the affected area. The architecture integrates LoRaWAN-based sensor communication and a cloud-based backend for real-time analysis and reporting. This demonstrated

the strong effectiveness of combining UAVs with camera systems and AI to enable rapid and accurate wildfire detection.

UAVs also allow for enhanced coverage and mobility compared to static systems. For instance, the work by Mowla et al. (2024) [6] demonstrated how integrating UAV-based aerial imagery with advanced CNNs improves accuracy even in challenging conditions such as dense foliage or smoke obstructions. Nevertheless, challenges persist. These include limited battery life, restricted flight duration, sensitivity to weather conditions, and the need for efficient edge AI models to process data without relying heavily on remote servers. Addressing these constraints is vital to enable long-term deployment and scalability of UAV systems for wildfire applications.

D. Data and Performance Evaluation

The effectiveness of AI-based wildfire detection models is intrinsically linked to the quality and diversity of datasets used for training and evaluation. A central challenge in the field is the limited availability of annotated, high-resolution wildfire imagery, especially from UAV platforms. To address this, recent studies have employed both custom and publicly available datasets collected from UAVs, satellites, closed-circuit television (CCTV) systems, and sensor networks, each of which contributes unique characteristics related to environmental conditions, fire phases, and image modalities due to different types of image acquisition.

Fig. 1 summarizes the main datasets referenced across key studies, including their image sources, types, and dataset sizes, corresponding to the columns *Collection Method*, *Type*, and *Number of Images*, respectively.

Name	Type	Collection Method	Number of Images	Reference	Dataset used in the Study
FLAME (2020)	RGB White-Hot Fusion Green-Hot	Drone	30,155 Fire and 17.855 No Fire	4	Only used RGB images. Used a subset of 5.615 images (not specified).
				8	Only used 6000 images: 3000 Fire and 3000 No Fire
				6	Used 35.600 images (not specified)
UAVs-FFDB (2024)	RGB	Drone	4 classes with 3.890 each. Two classes have Fire the other two do not	6	They used a drone to capture the images for the dataset
IIITDMJ_Smoke (2024)	RGB	Satellite MODIS	23.644 images	7	Covers different countries, years and different terrains
USTC_SmokeRS (2019/2024)	RGB	Satellite MODIS	4.059 images.	7	This dataset consists of a modification of the Ba et al. (2019)
Khan et al. (2019)	RGB	CCTV videos	72,012 images.	7	-
He et al. (2021)	RGB	CCTV videos	33.710 images.	7	-

Fig. 1: Summary of the main Datasets used in the literature

Among the most widely used datasets is FLAME (2020) [10], which comprises 48,010 aerial RGB and thermal images captured by drones, categorized into 30,155 “Fire” and 17,855 “No Fire” images. Studies such as Mowla et al. (2024) [6] and Rajalakshmi et al. (2023) [8] utilized subsets of FLAME (comprising up to 35,600 images), employing various train/validation/test splits (e.g., 85/10/5, 80/20). UAVs-FFDB (2024) [11], also utilized by Mowla et al. (2024) [6], comprises 15,560 drone-captured images across four classes, representing various times of day and fire intensities under real-world climate conditions. Mowla et al.’s AHMHCNN-mCBAM model [6] achieved 100% and >99% accuracy on UAVs-FFDB and FLAME, respectively, demonstrating exceptional performance in challenging scenarios.

Chaturvedi et al. (2024) [7] evaluated their hybrid convolution-transformer model across four datasets, which offer geographic and temporal diversity, covering different terrains and atmospheric conditions. Despite variations in image size and minimal smoke presence (as low as 2% of the image), their model achieved 93.90% accuracy on the USTC_SmokeRS dataset and over 99% accuracy on the other datasets.

Additionally, datasets such as those by Khan et al. (2019) and He et al. (2021), which contain 72,012 and 33,710 images, respectively, extracted from CCTV footage, are used to simulate real-time surveillance contexts. Though less common in UAV-based detection, they contribute to model training in dense, structured environments.

Despite promising results, generalization remains a significant concern. Many models experience performance drops when tested on unseen datasets with different conditions, such as, lighting, terrain, and vegetation types. This highlights the need for robust cross-dataset validation and broader benchmarks.

Performance is typically measured using metrics such as accuracy, precision, recall, F1-score, False Detection Rate (FDR), and Error Warning Rate (EWR). Increasingly, model interpretability and computational efficiency are also emphasized, especially for deployment in real-time embedded systems onboard UAVs, where processing power is constrained.

E. Limitations and Research Gaps

While AI-driven wildfire detection systems have demonstrated substantial progress, several limitations remain that hinder their full potential and operational effectiveness in real-world environments. One of the most persistent problems is the lack of model generalization across datasets and environmental conditions. As observed in the works of Mowla et al. [6] and Chaturvedi et al. [7], models trained on one dataset may underperform when tested on different types of terrain,

vegetation, lighting conditions, or image resolutions. This reduces the reliability of AI systems in dynamic, unseen scenarios.

Another critical gap lies in the reduction of false positives, particularly under challenging conditions such as fog, sunlight filtering through trees, or smoke from non-fire sources. Early efforts to address this problem, such as the intelligent verification system proposed by Arrue et al. [9], remain relevant, as modern systems continue to face high false alarm rates, especially in infrared-based detection.

In addition, computational limitations constrain real-time implementation, especially on low-power edge devices and UAV platforms. Although lightweight architectures such as the one proposed by Chaturvedi et al. [7] offer promising directions, balancing accuracy, model size, and inference speed remains a challenge.

Furthermore, limited and non-standardized datasets pose a major obstacle. Many studies rely on private datasets or simulations, making it difficult to compare models objectively. There is a growing call for the development of large, diverse, and publicly available UAV-based wildfire datasets with standardized annotation protocols [3].

Lastly, integration challenges between AI algorithms and UAV platforms, such as sensor fusion, real-time navigation, and regulatory compliance, must be addressed to enable scalable field deployment [3].

These limitations, in particular poor generalization across datasets, a lack of dataset standardization, and limited robustness in real-world conditions, underscore the need for more adaptable AI models. This work tackles these issues by evaluating cross-dataset generalization, applying data augmentation, and combining diverse UAV datasets to improve model robustness in unseen environments.

IV. Datasets and Data Acquisition

The effectiveness of AI-based models depends heavily on the quality and suitability of the datasets used for training and evaluation. FLAME, UAVs-FFDB, and FireMan-UAV-RGBT datasets were selected for the development of this project. Each dataset offers unique features and data formats captured by UAVs, which are suitable for fire classification, segmentation, and localization tasks.

A. Dataset Selection Criteria

The selection of datasets for this project was based on three key criteria: public availability, domain relevance, and diversity in environmental and visual conditions. These criteria ensure that the models are trained and evaluated on data that reflects real-world UAV-based fire detection scenarios.

The FLAME and UAVs-FFDB datasets were selected as primary sources due to their high-resolution annotations and imagery captured via UAVs, aligning with the intended future application of the developed models in real UAV platforms. As shown in Fig. 1, FLAME is a widely adopted benchmark dataset across multiple studies, confirming its robustness, relevance, and comparability in wildfire detection research. Meanwhile, UAVs-FFDB offers a well-structured collection of images representing both fire and no-fire conditions across different phases of the day, including evening and pre-evening scenarios, providing functional diversity to improve model generalization under variable lighting and atmospheric conditions.

In addition to these, the FireMan-UAV-RGBT dataset was also selected. Although it was not mentioned in the reviewed State of the Art, it was discovered during exploratory research and included for its complementary characteristics: it contains fire scenarios captured from angles and viewpoints not present in the other two datasets. This spatial diversity is expected to enhance the robustness of the trained model when deployed in heterogeneous and realistic fire-monitoring environments.

B. FLAME Dataset (2020)

The FLAME (Fire Luminosity Airborn-based Machine Learning Evaluation) dataset was collected during pile burn operations in a ponderosa pine forest near Flagstaff, Arizona. It includes a diverse collection of aerial images captured by drones equipped with RGB cameras and thermal sensors, such as DJI Matrice 200 and DJI Phantom 3. The dataset includes videos and frames of those videos in RGB and thermal imaging pallets (e.g., WhteHot, GreenHot, Fusion), as shown in Fig. 2. [12]

The FLAME dataset supports two main computer vision tasks: binary classification (fire vs no-fire) and pixel-wise segmentation of fire regions. It includes over 39,000 annotated frames for classification and more than 2,000 high-resolution frames for segmentation, accompanied by manually labeled masks. The dataset is publicly available via IEEE Dataport [10].

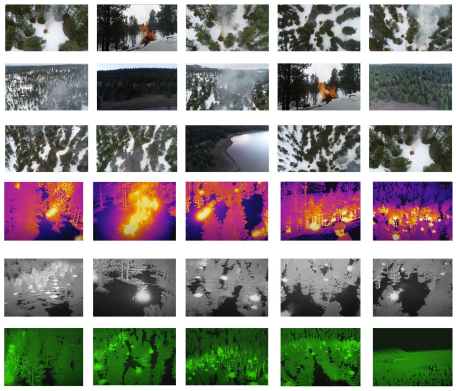

Fig. 2: Image examples of the FLAME dataset

C. UAVs-FFDB Dataset (2024)

The UAVs-FFDB (Unmanned Aerial Vehicles - Forest Fire Detection Database) dataset was developed at Alparslan Türkeş Science and Technology University, Turkey. It consists of 1,653 high-resolution RGB images captured using a Raspberry Pi Camera V2 mounted on an S500 quadrotor UAV. These images were captured in various environmental conditions and light levels, thereby promoting diversity in the dataset. In addition to the captured images, the dataset creators developed a set of augmented images through geometric transformations, including rotation, flipping, zooming, and shearing [13]. All fire-related images are divided into four classes: pre-evening and evening forest conditions and pre-evening and evening fire incidents. Both the original and augmented images were annotated using Makesense.ai [14] in XML format, which allows the dataset to be used for classification and object detection tasks, making it particularly valuable for developing generalizable AI models.

The dataset features images similar to those presented in Fig. 3 and is available on Mendeley Data [11].

Fig. 3: Image examples of the UAVs-FFDB dataset

D. FireMan-UAV-RGBT (2024)

The FireMan-UAV-RGBT dataset is a recent multimodal resource designed for wildfire detection using UAV-based RGB and thermal imagery. For this project, only the binary classification subset was considered. This subset includes annotated images labeled as either “Fire” or “No_Fire”, derived from drone videos recorded during controlled burn operations in Finland between 2022 and 2023. In Fig. 4, examples of images from the dataset are shown.

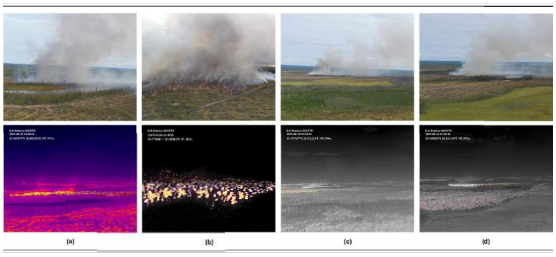

Fig. 4: Image examples of the FireMan-UAV-RGBT dataset

The dataset was acquired using DJI Matrice 30T and Matrice 300 UAVs equipped with Zenmuse H20T and H20N cameras, which captured both RGB and thermal video data. From these videos, frames were extracted and manually or semi-automatically annotated, with bounding boxes marking regions containing visible fire activity. Annotations are provided in YOLO format, ensuring compatibility with common object detection frameworks.

The FireMan-UAV-RGBT binary dataset was selected due to its realism, multimodal acquisition strategy, and precise labeling methodology. It is particularly suitable for training convolutional models on the fundamental task of distinguishing fire from no-fire instances, a necessary step before advancing to more complex tasks such as smoke detection or fire segmentation. This dataset is available in Zenodo [15].

E. Preprocessing

Although detailed preprocessing steps are part of the implementation phase, several recommendations from dataset creators and prior studies were followed to ensure data compatibility and improve model performance. Specifically, the authors of the FLAME and UAVs-FFDB datasets recommend resizing images to 254 X 254 or 256 X 256 pixels, respectively, and applying normalization to scale pixel values to the [0, 1] range. These recommendations were both adopted in this project to enhance training stability and convergence. Additionally,

data augmentation techniques suggested in the literature, including horizontal flipping, rotation, zooming, and contrast adjustments, were also considered. However, augmentation was only applied during a specific phase of the project focused on evaluating its impact on model generalization.

V. Methodology

This section describes the project's methodological framework, which aims to assess and improve the generalization abilities of DL models in aerial wildfire detection. The approach focuses on reproducing and adapting two state-of-the-art CNNs, instead of proposing novel architectures, and testing their ability to perform consistently across diverse datasets.

The methodology is organized into multiple stages, beginning with baseline reproduction to validate the implemented models, followed by cross-dataset evaluation to assess generalization. Further steps include the application of data augmentation strategies and dataset merging techniques to investigate their impact on model robustness. All experimental procedures were implemented in Python using TensorFlow and executed within GPU-enabled Jupyter Notebooks.

The following subsections describe the selected models, how datasets were used and divided, the development workflow, the computational environment, the training settings, and the evaluation metrics employed in the study.

A. Selected AI Models

This project evaluates two high-performing CNNs, chosen for their strong results in previous research and complementary architectural features: a lightweight hierarchical model by Ramadan et al. [4], and the AHMHCNN-mCBAM by Mowla et al. [6], which incorporates attention mechanisms.

Rather than introducing new architectures, the project faithfully reproduces these models based on published descriptions to assess whether their internal design choices, such as hierarchical structures or attention modules, enhance generalization across diverse UAV-based wildfire datasets.

Model 1 - Hierarchical CNN

The hierarchical CNN proposed by Ramadan et al. [4] was selected for its simplicity, fast execution, and proven applicability in UAV deployments. Although it does not include advanced attention mechanisms, its architecture is optimized for low-latency processing on drones and includes:

- **Convolutional Feature Pyramid:** A sequence of hierarchical convolutional layers designed to progressively extract low- to high-level features.
- **Batch Normalization and ReLU Activations:** Included after each convolution to stabilize training and improve performance.
- **Global Average Pooling:** Used to reduce dimensionality and avoid overfitting by minimizing the number of trainable parameters.
- **Compact Fully Connected Head:** A small dense block that performs final classification, leading into a softmax output layer.

The FLAME dataset was used to train this model, which demonstrated success in real-time onboard fire detection with an accuracy of over 99%, serving as a benchmark fast quick and effective deployment. Henceforth, this model will be referred to as Model 1.

Model 2 - AHMHCNN-mCBAM

The AHMHCNN-mCBAM proposed by Mowla et al [6], integrates a hierarchical, multi-headed CNN with a Modified Convolutional Block Attention Module (mCBAM), which enhances the model's focus on fire-specific patterns in aerial images. Its multi-scale feature extraction and attention mechanism are designed to improve recognition of localized and context-sensitive fire cues, which are critical for real-world variability. The model has shown excellent performance on standard datasets such as UAVs-FFDB and FLAME, achieving accuracies above 99% and superior metrics in fire/no-fire classification tasks. The model is built upon a modular design of four key components:

- **Multi-Scale Feature Extraction:** The model employs multiple parallel convolutional blocks with varied kernel sizes (e.g., 3×3, 5×5, 7×7) to capture both local and global patterns in the input images.
- **Modified CBAM (mCBAM):** Each convolutional branch is augmented with a channel and spatial attention block, enhanced with normalization and max/average pooling to focus on fire-relevant visual cues.
- **Adaptive Pooling and Dense Layers:** The feature maps from different branches are concatenated, pooled, and passed through fully connected layers.
- **Softmax Output:** The final layer employs a softmax activation function for binary classification (fire vs. no-fire), trained using categorical cross-

entropy.

This architecture is specifically designed to enhance the model's ability to attend to semantically relevant areas within UAV images, potentially improving generalization when applied to different datasets. Henceforth, this model will be referred to as Model 2.

B. Sub-Datasets and Splits

It is common in the literature for authors to cite publicly available datasets but ultimately rely on only a limited portion of the data without clarification, affecting transparency and reproducibility. To ensure this does not occur in this report, this subsection is dedicated to describing how the data was selected and divided.

Throughout this work, whenever we state that a model was trained with the FLAME, UAVS-FFDB, or FireMan-UAV-RGBT datasets, we are explicitly referring to the portions and data splits described below.

The **FLAME** dataset is organized into multiple folders by the dataset creators. Of particular relevance are Repository 7, intended for training and validation, and Repository 8, intended for testing. In line with the procedure described by Ramadan et al. [4] (which we replicate here), both repositories are merged to form a new balanced subset comprising 5,615 images with 2,854 fire and 2,761 no-fire examples. We refer to this subset as *FLAME_subset*. The original folders are retained as *FLAME_train* (Repository 7) and *FLAME_test* (Repository 8) and are used later for generalization evaluation. The *FLAME_subset* was randomly split into 80% training, 10% validation, and 10% test, regardless of its use in Model 1 or 2.

The **UAVS-FFDB** dataset is divided into two distinct parts: Raw Images and Augmented Images. Following a recommendation received directly from one of the dataset's creators, we used *UAVS_RawImages* (all original, unaltered images in the dataset) for model training and validation, with a 90%/10% split, respectively. The complete set of *UAVS_AugImages* (all the augmented images in the dataset) was used as the dedicated test split, as per the same recommendation. Both *UAVS_RawImages* and *UAVS_AugImages* represent the entirety of the UAVS-FFDB dataset, with no additional filtering applied.

The **FireMan-UAV-RGBT** dataset is distributed with predefined partitions, originating the following datasets: *FireMan_train*, *FireMan_val*, and *FireMan_test*. They correspond to the original splits provided by the dataset authors and are used as-is throughout this study. Collectively, they represent the complete dataset.

TABLE I shows the used data splits and TABLE II summarises all used datasets, listing the number of fire, no-fire, and total images in each.

TABLE I: Train/Val/Test Splits for Each Dataset

Dataset	Split Train	Split Val	Split Test
FLAME	80% FLAME_subset 4492 images 2283 Fire; 2209 No_Fire	10% FLAME_subset 562 images 286 Fire; 276 No_Fire	10% FLAME_subset 561 images 285 Fire; 276 No_Fire
UAVs-FFDB	90% UAVS_RawImages 1478 images 1030 Fire; 448 No_Fire	10% UAVS_RawImages 165 images 115 Fire; 50 No_Fire	UAVS_AugImages 15560 images 7780 Fire; 7780 No_Fire
FireMan-UAV-RGBT	FireMan_train 5313 images 4883 Fire; 430 No_Fire	FireMan_val 611 images 588 Fire; 23 No_Fire	FireMan_test 589 images 556 Fire; 33 No_Fire

TABLE II: Distribution of the Datasets

Dataset	Number of Images	Images with Fire	Images without Fire
FLAME subset	5615	2854	2761
FLAME_train	39375	25018	14357
FLAME_test	8617	5137	3480
UAVS_RawImages	1643	1145	498
UAVS_AugImages	15560	7780	7780
FireMan_train	5313	4883	430
FireMan_val	611	588	23
FireMan_test	589	556	33

In addition to evaluating model performance on the test split, we conducted a series of cross-dataset tests to assess how well the models generalize to previously unseen data sources. This is particularly important in fire detection scenarios, where environmental conditions, image characteristics, and sensor viewpoints can vary significantly from one dataset to another.

To this end, the following datasets were used for generalization testing: *FLAME_train*, *FLAME_test*, *UAVS_RawImages*, *UAVS_AugImages* and *FireMan_test*.

These evaluations are designed to measure model robustness across variations in the dataset domain, including differences in terrain, lighting, image resolution, and acquisition devices. By comparing performance on these independent datasets, we aim to understand better a model's ability to generalize beyond its training distribution.

C. Development Worflow

The methodological workflow of this project is divided into four distinct phases, each with a specific objective and evaluation. Beyond reproducing results from previous studies, each phase is designed to assess a particular aspect of the model's generalization ability, ranging from baseline replication to its performance on unseen data and the impact of data augmentation and dataset combination strategies.

Phase 1: Baseline Reproduction

The initial phase focuses on reproducing the two selected models, the hierarchical CNN and the AHMHCNN-mCBAM, using their original datasets as reported in the respective studies. For *Model 1*, we utilized the FLAME dataset, and for *Model 2*, we employed the UAVs-FFDB dataset, as explained in Section V-B. The goal is to ensure that the recreated implementations yield comparable performance metrics (e.g., accuracy, precision, recall), providing a reliable foundation for generalization testing, which is essential for validating the consistency of the implemented models and serves as a baseline for future comparisons.

Phase 2: Cross-Dataset Generalization

In this phase, the focus shifts to evaluating how well models trained on one dataset perform on unseen datasets. Each model is trained independently on the FLAME, UAVs-FFDB, and FireMan datasets and then tested on the generalization datasets, which were described in Section V-B. This cross-dataset evaluation reveals the extent to which each model can adapt to new data distributions, which is an essential requirement for reliable performance in real-world wildfire detection scenarios

Phase 3: Data Augmentation for Generalization

Data augmentation plays a critical role in enhancing the generalization capabilities of CNNs, particularly in image classification tasks. By artificially increasing the variability of the training data, it reduces overfitting and improves performance on unseen datasets. The paper by Mowla et al. [6] explicitly states that augmentations "enhance performance on unseen data" and improve accuracy in real-world fire detection scenarios. The documentation of the FLAME dataset [12] also recommends applying transformations, such as rotation and flipping, to enhance the model's adaptability.

Considering these insights, data augmentation suggested by Mowla et al. [6] was incorporated into the training process using the *ImageDataGenerator* class from TensorFlow, which enables dynamic image transformations during batch generation. TABLE III describes the configurations applied:

TABLE III: Data Augmentation Applied

Technique	Parameters
Rotation	rotation_range=160
Width shift	width_shift_range=0.2
Height shift	height_shift_range=0.2
Shear	shear_range=0.2
Zoom	zoom_range=0.4
Horizontal flip	horizontal_flip=True
Fill mode	fill_mode='reflect', cval=125

This configuration features extensive rotation, zoom, shear, and brightness adjustments to simulate various real-world scenarios. As recommended by Mowla et al. (2024) [6], this augmentation setup aims to boost the model's ability to recognize wildfire patterns across different environments.

Phase 4: Dataset Combination Strategy

This phase investigates whether combining datasets during training improves generalization. Images from FLAME, UAVs-FFDB, and FireMan were resized (based on model input requirements), normalized, and merged into balanced training sets. The hypothesis is that exposing the model to greater visual diversity during training will enable it to adapt more effectively to unseen environments.

The guiding principle for dataset combination was based on the number of available images in each dataset. Unlike traditional approaches that prioritise class balance (i.e., equal distribution of "fire" and "no-fire" samples), this strategy focused on balancing the dataset sizes across sources, while allowing the natural distribution of classes to remain unchanged. For every combination, the dataset with the smallest number of available training and validation images was used to define the maximum number of samples that each of the other datasets could contribute. Any exceeding images from the larger datasets were not discarded but instead allocated exclusively to the test split of the combined set.

Each dataset contributed to the combined sets as follows:

- The *FLAME_subset* was partitioned such that 90% of its images are used for the training and validation splits, while the remaining 10%, originally intended as a validation split, was exclusively assigned to the test split in all combined configurations. The full *FLAME_test* directory was also reserved solely for the test split.

- On UAVS-FFDB dataset only the original, unaltered images from the *Raw_Images* folder were considered for training and validation purposes. The entire *UAVS_AugImages* set was excluded from training and used only as part of the test split.
- The *FireMan_train* and *FireMan_val* sets were merged and treated as a unified source of data for training and validation. The *FireMan_test* set was always assigned to the test split and never included in training.

TABLE IV shows the final distributions of the combined datasets:

TABLE IV: Combined Datasets Distribution

Combination	Total Images	Fire	No Fire	FLAME	UAVS-FFDB	FireMan
FLAME and UAVS-FFDB train	3286	1995	1291	1643	1643	-
FLAME and UAVS-FFDB test	19532	9784	9784	3972	15560	-
FLAME and FireMan train	10668	7756	2910	5053	-	5053
FLAME and FireMan test	1460	1123	337	562	-	898
UAVS-FFDB and FireMan train	3260	2661	625	-	1643	1643
UAVS-FFDB and FireMan test	20430	12291	8139	-	15560	4870
All datasets	4929	3511	1418	1643	1643	1643
All datasets	24402	14295	10107	3972	15560	4870

D. Tools and Environment

The implementation of this project was conducted using Python 3.10.14 in a Jupyter Notebook environment. All code was executed with GPU acceleration enabled, using configurations optimized for memory efficiency, which leveraged TensorFlow's *cuda_malloc_async* and dynamic memory growth features.

The main libraries used are summarized in TABLE V below:

TABLE V: Tools and Versions Used

Tool	Version
Python	3.10.14
TensorFlow	2.13.0
NumPy	1.23.5
Pandas	1.5.3
Matplotlib	3.7.1

E. Training Settings

Understanding the training settings is crucial, as they have a significant impact on a model's performance, generalization ability, and reproducibility. These settings influence how well a model learns from data and whether its results can be reliably replicated or compared. Transparent reporting of training configurations also helps ensure fair evaluation across different models. The training of both Model 1 and Model 2 was conducted according to the settings detailed in TABLE VI. As shown in the table, the majority of parameters are similar, reflecting the recommended values provided by the original authors of each model. Nonetheless, some differences can be observed in specific configurations, such as the *Input Dimensions*, and the *Callbacks*.

TABLE VI: Training Hyperparameters

Hyperparameters	Model 1	Model 2
Number of epochs	50	50
Batch Size	16	16
Activation	Softmax	Softmax
Optimizer	Adam	Adam
Loss Function	Categorical Crossentropy	Categorical Crossentropy
Input Dimensions	224 X 224 pixels	256 X 256 pixels
Callbacks	ModelCheckpoint	ModelCheckpoint Lr Schedule

The **batch size** value was selected based on GPU memory limitations.

Softmax was selected as the **activation function** since both studies of the chosen architectures for the baseline reproduction use it.

The Adam **optimizer** is known for its adaptive learning rates and stability in DL applications and is recommended by both Mowla et al. [6] and Shamsoshoara et al. [12], as such it was also our choice.

For the **loss function**, categorical cross-entropy was chosen, as suggested by Mowla et al. [6] that recommends the use of this loss functions.

The **input dimensions** vary accordingly to the used model, as that is the way the baseline architectures were done.

Regarding the **callbacks**, in both models, to save the model weights corresponding to the lowest validation loss during training, the ModelCheckpoint was added, ensuring that the best-performing version of the model on unseen data is retained. The learning rate scheduling strategy for Model 2 is the one described by Mowla et al. [6] and is used to adjust the Learning Rate (Lr) during training to improve convergence:

$$\text{Lr schedule:} \begin{cases} 1 \times 10^{-4}, & \text{for epoch} < 15 \\ 1 \times 10^{-5}, & \text{for } 15 \leq \text{epoch} < 25 \\ 1 \times 10^{-6}, & \text{for epoch} \geq 25 \end{cases}$$

All training procedures were performed with GPU acceleration enabled to ensure computational efficiency.

F. Evaluation Metrics

Evaluation metrics are essential for determining a model's efficiency, and selecting the right metrics is

crucial, as they enable researchers to assess how well a model performs on a particular task.

To evaluate the performance of each model, we compute the following metrics derived from the confusion matrix values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The performance metrics and their corresponding formulas used for model evaluation are summarized in TABLE VII, such as Accuracy, Precision, Recall, F1-Score, Area Under the ROC Curve (AUC), Loss, FDR and EWR.

TABLE VII: Performance Metrics Formulas

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
AUC	`roc_auc_score` calculated with *sklearn.metric*
Loss	$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij})$
Fire Detection Rate (FDR)	Recall $\times$ 100
Error Warning Rate (EWR)	$\frac{FP+FN}{TP+TN+FP+FN} \times 100$

These metrics were computed programmatically after model inference, and confusion matrices were used to visualize classification outcomes.

VI. Results

Comparing results is essential to determine whether the intended objectives have been successfully achieved. By analyzing the performance of the models against the chosen evaluation metrics, it becomes possible to assess their effectiveness and identify any areas where they fall short. This comparison not only validates the training process but also provides insights into the strengths and limitations of each approach, guiding future improvements and informing conclusions.

The section presents and analyses the performance of the models across all phases of the study, using quantitative metrics to evaluate their generalization ability and effectiveness in wildfire detection.

As mentioned in Subsection V-E, the training process employed the ModelCheckpoint callback, which was used in this case to save the version of the model that achieved the lowest validation loss during training. As a result, two versions of each model were saved at the end of the training process: the *final_model*, representing the model at the last training epoch, and the *mc_model*, which corresponds to the model checkpoint saved at the point of best validation performance in terms of loss. For each of the four project phases and their respective evaluations, we report results based solely on the better-performing model between the *final_model* and the *mc_model*. That is, for any given test (e.g., Phase 1 using the FLAME dataset), the results will reflect either the *final_model* or the *mc_model*, but never a combination of the two.

A. Phase 1: Baseline Reproduction

In this subsection, we evaluate how well the models were recreated in comparison to the original architectures that served as their foundation. The objective is to compare key performance metrics between each recreated model and its corresponding reference to assess the fidelity of the replication process.

TABLE VIII compares the results of Model 1 and its inspiration, the Hierarchical CNN and TABLE IX compares Model 2 and its baseline, the AHMHCNN-mCBAM model. The confusion matrices for each model are presented in Appendix A.

TABLE VIII: Model 1 vs Hierarchical CNN on FLAME Dataset

Metric	Model 1	Hierarchical CNN
Dataset	FLAME	FLAME
Loss	0,0427	–
Accuracy	98,93	99,46
Precision	98,27	99,64
Recall	99,65	99,29
F1–Score	98,95	99,46
AUC	99,86	–
FDR	99,69	–
EWR	1,07	–

To determine whether the model replication was successful, the performance of Model 1 was compared to the original Hierarchical CNN. Despite a few minor variations, the results show that the two models are generally very similar.

Model 1 has an Accuracy of 98,93%, which is slightly less than the Hierarchical CNN's reported Accuracy of 99,46%. Additionally, the recreated model's Precision is marginally lower (98,27% vs. 99,64%), suggesting a slight rise in false positives. On the other hand, Model 1 has a slightly better Recall (99,65% vs. 99,29%), indicating a marginally better capacity to identify true

positives. Once again indicating a slight decline in overall performance, the recreated model's F1-score, which weighs Precision and Recall, was 98,95% compared to 99,46% for the original model.

Notably, the recreated model's Loss value was low (0,0427%), and its AUC stayed high (99.86%), both of which are signs of excellent overall performance.

The overall performance metrics are very similar to those of the original Hierarchical CNN despite some slight decreases in Accuracy, Precision, and F1-score. The model's high Accuracy and outstanding Recall indicate that the main traits and behaviors of the original model were successfully replicated.

TABLE IX: Model 2 vs AHMHCNN-mCBAM on UAVS-FFDB Dataset

Metric	Model 2	AHMHCNN-mCBAM
Dataset	UAVS-FFDB	UAVS-FFDB
Loss	0,0247	–
Accuracy	99,41	100,00
Precision	100,00	100,00
Recall	98,82	100,00
F1–Score	99,41	100,00
AUC	99,99	–
FDR	98,82	100,00
EWR	0,59	0,00

Model 2 was compared with the original AHMHCNN-mCBAM model to evaluate the fidelity of the replication process. According to the analysis, there are only slight variations between the recreated model and the original in several key performance metrics.

Both models achieve perfect Precision (100,00%), indicating no false positives, and Model 2 maintains an exceptionally high AUC of 99,99%, suggesting outstanding discriminatory power. However, Model 2 shows a modest decrease in Accuracy (99,41% compared to 100,00%), as well as in Recall (98,82% vs 100,00%), which slightly impacts the overall F1-score (99,41% vs 100,00%)

In conclusion, Model 2's replication shows an excellent approximation of the original AHMHCNN-mCBAM. Although there are some slight variations, especially in Recall and FDR, the model successfully maintains the essential features and excellent performance of the original. All things considered, this is a very successful and accurate replication.

B. Phase 2: Cross-Dataset Generalization

In Phase 2, the models' generalization across datasets, that is, their performance on image data different from the source used for training, is assessed. Determining whether these models can be effectively implemented in real-world settings, where UAV-captured imagery varies significantly in terms of lighting, terrain, resolution, and sensor type, requires such an evaluation. To achieve this, each model was trained separately on the FLAME, UAVs-FFDB, and FireMan-UAV-RGBT datasets. It was then evaluated on the remaining datasets to determine its robustness in cross-domain scenarios.

Generalization Capacity with FLAME dataset

In this test, both Models 1 and 2 were trained on the FLAME dataset as described in Subsection V-B and its results can be seen in TABLE X and in TABLE XI. The confusion matrix components for each model are presented in Appendix B.

TABLE X: Model 1 - FLAME Generalization

Metric	Model 1: FLAME					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0427	0,0156	0,0536	5,1866	3,8150	8,2291
Accuracy	98,93	99,59	98,17	22,82	39,33	15,79
Precision	98,27	99,84	98,03	28,72	33,60	100,00
Recall	99,65	99,51	98,91	7,25	21,85	10,79
F1–Score	98,95	99,67	98,47	11,58	26,48	19,48
AUC	99,86	99,98	99,83	33,54	32,53	94,13
FDR	99,65	99,51	98,91	7,25	21,85	10,79
EWR	1,07	0,41	1,83	77,18	60,67	84,21

Model 1 performs exceptionally well on its test split and also on semi-observed data (FLAME_train and FLAME_test, showing strong F1-scores, near-perfect AUC, and high Accuracy (99,59% and 98,17%), confirming a robust source-domain fit.

However, the model's performance drastically deteriorates when used on unseen datasets. On the *UAVS_RawImages* dataset, Recall (7,25%) is incredibly low, Accuracy drops to 22,82%, and the F1-score drops to 11,58%. Although there is a minor improvement with the *UAVS_AugImages* dataset, the overall performance remains subpar. More concerningly, the evaluation of *FireMan_test* yields a high EWR of 84,21%, along with a low Accuracy of 15,79% and a Recall of 10,79%.

Notebly, the model achieves a very high Precision (100,00%) and an AUC of 94,13% on the *FireMan_test* dataset. However, this is misleading when viewed alongside the confusion matrix, which reveals that the model misses the vast majority of true positives. The high Precision derives from the absence of false positives, while the elevated AUC likely reflects a decent ranking of probabilities rather than effective classification. This

suggests that the model is overly conservative, only predicting positives when it is highly confident.

TABLE XI: Model 2 - FLAME Generalization

Metric	Model 2: FLAME					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0320	0,0126	0,0467	2,7881	1,6296	5,8582
Accuracy	99,47	99,66	98,53	40,47	51,31	14,09
Precision	99,65	99,85	99,41	73,52	53,62	100,00
Recall	99,30	99,62	98,11	22,79	19,43	8,99
F1–Score	99,47	99,73	98,76	34,80	28,53	16,50
AUC	99,75	99,98	99,86	57,17	64,03	20,25
FDR	99,30	99,62	98,11	22,79	19,43	8,99
EWR	0,53	0,34	1,47	59,53	48,69	85,91

With an Accuracy of 99,66% on the *FLAME_train* and 98,53% on the *FLAME_test*, Model 2 demonstrates excellent performance on semi-observed data with consistently high F1-scores and AUC values across these sets.

Model 2 outperforms Model 1 in terms of generalizing to new datasets despite not necessarily having good results. With a Precision of 73,52%, its Accuracy increases to 40,47% and its F1-score to 34.80% on *UAVS_RawImages* dataset. Additionally, the model performs better on the *UAVS_AugImages* dataset, attaining the highest external Accuracy (51,31%) of all tested scenarios.

Although Model 2 also achieves perfect Precision (100,00%) on *FireMan_test*, its AUC is much lower (20,25%). This suggests that, unlike Model 1, it struggles to distinguish between positive and negative cases based on probability scores. The confusion matrix shows that the model predicts very few positives and fails to rank most true positives above negatives, leading to poor overall separability and a low AUC.

Generalization Capacity with UAVS-FFDB dataset

As explained in Subsection V-B, the UAVS-FFDB dataset was used to train both Models 1 and 2 for this test, and the generalization results are shown in TABLE XII and XIII, respectivelly. The confusion matrix components for each model are presented in Appendix C.

TABLE XII: Model 1 - UAVS Generalization

Metric	Model 1: UAVS					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,3679	12,9512	8,8205	0,0385	0,3679	5,7776
Accuracy	96,81	18,66	42,72	99,63	96,81	55,69
Precision	96,07	1,40	69,67	99,91	96,07	90,86
Recall	97,62	0,40	6,93	99,56	97,62	58,99
F1–Score	96,84	0,63	12,61	99,74	96,84	71,54
AUC	98,74	6,79	48,53	99,86	98,74	20,95
FDR	97,62	0,40	6,93	99,56	97,62	58,99
EWR	3,19	81,34	57,28	0,37	3,19	44,31

Model 1 exhibits limited generalization capabilities on unseen data, particularly on *FLAME_train*, where it achieves only 18,66% Accuracy, 1,40% Precision, and a low Recall of 0,40%, resulting in an F1-score of 0,63%. On the *FLAME_test* dataset, the model performs slightly better. With a Precision of 69,67%, it achieves 42,72% Accuracy but struggles with Recall (6,93%) due to a high number of false negatives. This results in an F1-score of 12,61%, highlighting the model's difficulty in capturing positive cases. Performance improves significantly on *FireMan_test*. Here, the model achieves 55,69% Accuracy, 90,86% Precision, and 58,99% Recall, yielding a respectable F1-score of 71,54%. However, it struggles to effectively distinguish between positive and negative classes. The absence of true negatives and the large number of both false positives and false negatives indicate a lack of reliable class separation.

TABLE XIII: Model 2 - UAVS Generalization

Metric	Model 2: UAVS					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0247	9,0810	8,3571	0,0015	0,0247	12,9773
Accuracy	99,41	36,21	40,28	100,00	99,41	5,60
Precision	100,00	17,53	0,00	100,00	100,00	0,00
Recall	98,82	0,11	0,00	100,00	98,82	0,00
F1–Score	99,41	0,21	0,00	100,00	99,41	0,00
AUC	99,99	17,32	31,02	100,00	99,99	100,00
FDR	98,82	0,11	0,00	100,00	98,82	0,00
EWR	0,59	63,79	59,72	0,00	0,59	94,40

Model 2 shows critical generalization failure. On *FLAME_train*, despite achieving an Accuracy of 36,21%, the model exhibits an extremely low Recall of 0,11% and an F1-score of only 0,21%, indicating a near-total failure to identify positive samples. On *FLAME_test*, performance deteriorates completely, with zero true positives and 5137 false negatives, resulting in Precision, Recall, and F1-score all at 0,00%. The same occurs on *FireMan_test*, where the model again predicts only the negative class, missing all 556 positives. These results

reflect a severe class imbalance bias and a model that is overfitted and unable to adapt to any unseen domain.

Generalization Capacity with FireMan dataset

TABLE XIV and TABLE XV show the results of the generalization tests of Models 1 and 2 trained with the FireMan dataset. The confusion matrix components for each model are presented in Appendix D.

TABLE XIV: Model 1 - FireMan Generalization

Metric	Model 1: FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0032	2,1142	4,7086	3,9863	6,6834	0,0032
Accuracy	100,00	79,50	60,64	62,99	37,64	100,00
Precision	100,00	75,68	60,31	67,94	42,69	100,00
Recall	100,00	99,82	99,38	88,82	72,20	100,00
F1–Score	100,00	86,09	75,06	76,99	53,66	100 ,00
AUC	100,00	78,43	54,57	28,03	26,37	100,00
FDR	100,00	99,82	99,38	88,82	72,20	100,00
EWR	0,00	20,5	39,36	37,01	62,36	0,00

Model 1 shows varying degrees of generalization to unseen datasets. On the *FLAME_train* dataset, it achieves a high Accuracy of 79,50%, F1-score of 86,09%, and Recall of 99,82%, indicating strong sensitivity. The confusion matrix confirms this high Recall, though the relatively high number of false positives impacts precision (75,68%). On *FLAME_test*, Accuracy drops to 60,64% and F1-score to 75,06%, with a Precision of 60,31%. Still, Recall remains high at 99,38%. Despite a decrease in overall performance, the model reliably identifies positive cases, although with some failures. For *UAVS_RawImages*, Accuracy is 62,99%, F1-score 76,99%, and Precision 67,94% with a solid Recall of 88,82%. Finally, on *UAVS_AugImages*, performance declines notably, with an Accuracy of 37,64%, an F1-score of 53,66%, and a Precision of 42,69%. The confusion matrix highlights the difficulty, especially with false positives.

TABLE XV: Model 2 - FireMan Generalization

Metric	Model 2: FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,2395	2,3875	1,5492	1,6801	2,2056	0,2395
Accuracy	92,02	54,19	61,77	72,31	59,06	92,02
Precision	100,00	60,27	63,16	75,56	56,39	100,00
Recall	91,55	81,92	86,10	89,08	79,95	91,55
F1–Score	95,59	69,44	72,87	81,76	66,13	95,59
AUC	100,00	14,56	57,80	45,14	47,80	100,00
FDR	91,55	81,92	86,10	89,08	79,95	91,55
EWR	7,98	45,81	38,23	27,69	40,94	7,98

Model 2 also shows varied performance across external datasets. On *FLAME_train*, the Accuracy is lower than that of Model 1, at 54,19%, with a Recall of 81,92% and a Precision of 60,27%. Its confusion matrix reinforces these values with 13512 false positives, 20494 true positives and 4524 false negatives.

On *FLAME_test*, results improve: Accuracy reaches 61,77%, the F1-score is 72,87%, and Recall is 86,10%. The confusion matrix indicates strong Recall despite high false positive rates, which suppresses Precision (63,16%). In *UAVS_RawImages*, the model achieves 72,31% Accuracy, an F1-score of 81,76%, and a Recall of 89,08%. For *UAVS_AugImages*, performance degrades, although not as severely: Accuracy is 59,06%, F1-score is 66,13%, and Recall is 79,95%.

Results Overview

The results of this phase demonstrate that strong generalization is not always correlated with high performance on the training dataset. Models developed using the FLAME dataset struggled with external datasets, with sharp declines in Accuracy and Recall. In a similar vein, models trained on UAVs-FFDB failed to generalize effectively on the FLAME and FireMan datasets despite achieving near-perfect results within their own domain. This strengthens a tendency towards overfitting and limited adaptability when trained on visually homogeneous data.

On the other hand, across all test sets, models trained on the FireMan dataset performed the most evenly and consistently. This suggests that FireMan's images offer a more solid basis for learning representations that can be applied to other contexts. Overall, these results reinforce the ongoing challenge of generalizing across datasets in UAV-based wildfire detection and the importance of selecting representative and varied training data for creating robust AI models suitable for practical applications.

C. Phase 3: Data Augmentation for Generalization

This phase investigates whether data augmentation improves the generalization of Model 1 and 2, since models trained on limited or homogeneous datasets often struggle with unseen data. It evaluates whether applying transformations, as described in Subsection V-C, helps the models better recognize fire patterns across varied environments.

Generalization Capacity with FLAME dataset with Data Augmentation

The results of the generalization tests of Models 1 and 2 trained with the FLAME dataset with data augmentation are displayed in TABLE XVI and XVII, respectively. The confusion matrix components for each model are presented in Appendix E.

TABLE XVI: Model 1 - FLAME Generalization with Data Augmentation

Metric	Model 1: FLAME (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,1199	0,0926	0,3220	3,2335	2,9363	7,3431
Accuracy	93,77	95,71	82,85	59,22	50,17	9,17
Precision	90,61	94,55	81,82	86,37	50,30	100,00
Recall	97,90	98,95	91,57	49,26	29,11	3,78
F1–Score	94,12	96,70	86,42	62,74	36,88	7,28
AUC	99,35	99,63	93,63	55,06	38,06	99,48
FDR	97,90	98,95	91,57	49,26	29,11	3,78
EWR	6,23	4,26	17,15	40,78	49,83	90,83

When tested on unseen datasets, Model 1, which was trained on the FLAME dataset, performs poorly in terms of generalization. With a comparatively high Precision of 86,37%, it demonstrated a conservative bias towards positive predictions on the *UAVS_RawImages* dataset, achieving an Accuracy of 59,22%, a Recall of 49,26%, and an F1-score of 62,74%. On *UAVS_AugImages*, however, performance significantly deteriorated, with F1-score falling to 36,88%, Recall to 29,11%, and Accuracy to 50,17%. The model's generalization completely failed on *FireMan_test*, achieving only 3,78% Recall, 9,17% Accuracy, and an F1-score of 7,28% despite having perfect precision of 100,00%. Poor adaptability to unknown data distributions is further indicated by the elevated EWR values (40,78%, 49,83%, and 90,83%) across all unseen datasets.

The use of augmentation decreased performance on *FireMan_test*, where Recall decreased from 10,79% to 3,78%, but marginally improved generalization on *UAVS_RawImages* and *UAVS_AugImages* when compared to the model without data augmentation (VI-B). As a result, data augmentation had conflicting results and provided little help in this instance.

TABLE XVII: Model 2 - FLAME Generalization with Data Augmentation

Metric	Model 2: FLAME (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0871	0,0410	0,2880	0,8753	1,0208	0,8799
Accuracy	97,69	97,55	89,51	56,73	59,42	83,87
Precision	98,23	98,72	89,57	77,47	58,86	93,74
Recall	97,20	97,41	93,6	53,45	62,57	88,85
F1–Score	97,72	98,06	91,38	63,26	60,66	91,23
AUC	99,64	99,84	95,05	74,53	76,94	25,47
FDR	97,20	97,41	93,26	53,45	62,57	88,85
EWR	2,31	2,45	10,49	43,27	40,58	16,13

When trained with data augmentation, Model 2 showed significantly better generalization on unseen datasets. The model demonstrated median adaptability on *UAVS_RawImages* dataset, achieving 56,73% Accuracy, 53,45% Recall, and a robust F1-score of 63,26%. With 59,42% Accuracy, 62,57% Recall, and an F1-score of 60,66%, the model continued to perform well on the *UAVS_AugImages* dataset. On the fireman dataset, the model demonstrated excellent generalization, achieving Accuracy of 83,87%, Recall of 88,85%, F1-score of 91,23%, and Precision of 93,74%.

This is a significant performance improvement over the version without augmentation (VI-B). For instance, Recall increased from 8,99% to 88,85% on *FireMan_test* and from 22,79% to 53,45% on *UAVS_RawImages*. EWR also significantly decreased across all external datasets. Thus, in Model 2, data augmentation significantly improved generalization.

Generalization Capacity with UAVS-FFDB dataset with Data Augmentation

TABLE XVIII and XIX show how well Model 1 and 2 generalize when trained on UAVS-FFDB with data augmentation, respectively. The confusion matrix components for each model are presented in Appendix E.

TABLE XVIII: Model 1 - UAVS-FFDB Generalization with Data Augmentation

Metric	Model 1: UAVS-FFDB (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0242	4,5953	5,8676	0,0333	0,0242	7,4245
Accuracy	99,46	29,13	34,52	99,27	99,46	16,47
Precision	98,98	34,83	42,28	99,30	98,98	74,62
Recall	99,95	13,24	26,92	99,65	99,95	17,45
F1–Score	99,46	19,19	32,90	99,48	99,46	28,28
AUC	100,00	43,20	31,99	99,97	100,00	11,23
FDR	99,95	13,24	26,92	99,65	99,95	17,45
EWR	0,54	70,87	65,48	0,73	0,54	83,53

Model 2, trained using data augmentation, demonstrated excellent performance on in-domain datasets, such as the *UAVS_RawImages* and Split Test (*UAVS_AugImages*), with Accuracy values of 99,27% and 99,46% respectively, and F1-scores also at 99,48% and 99,46%. This confirms that the model learned the training distribution well. However, its ability to generalize to unseen datasets was mixed. On the *FLAME_train* dataset, the model achieved low Accuracy (29,13%), Precision (34,83%), and Recall (13,24%), with a high EWR of 70,87%. On the *FLAME_test* dataset it showed marginal improvement, with an F1-score of 32,9% and EWR of 65,48%. On *FireMan_test*,

despite a high Precision of 74,62%, the Recall dropped significantly to 17,45%, leading to a modest F1-score of 28,28% and a very high EWR of 83,53%. These results suggest that while augmentation improved robustness in some unseen domains (particularly the FLAME dataset), the model struggled significantly the *FireMan_test* dataset.

TABLE XIX: Model 2 - UAVS-FFDB Generalization with Data Augmentation

Metric	Model 2: UAVS-FFDB (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0088	9,0587	8,2179	0,0108	0,0088	7,261
Accuracy	99,9	32,26	40,25	99,82	99,9	14,09
Precision	99,79	5,3	0,00	99,74	99,79	100,00
Recall	100,00	0,39	0,00	100,00	100,00	8,99
F1–Score	99,9	0,73	0,00	99,87	99,9	16,5
AUC	100,00	7,07	26,97	100,00	100,00	100,00
FDR	100,00	0,39	0,00	100,00	100,00	8,99

Model 2 exhibits near-zero generalization on both *FLAME_train* and *FLAME_test*, with extremely low performance across all evaluation metrics, including Recall, Precision, and F1-score. On the *FireMan_test* dataset, the model achieves a Precision of 100% but a Recall of only 8,99%, indicating a highly conservative behavior where almost no detections are made unless highly certain.

When compared to the non-augmented version of Model 2 (Subsubsection VI-B), the difference is subtle. Without augmentation, Recall and F1-score on the FLAME datasets and *FireMan_test* are effectively 0%, with Precision still at 0% on key datasets. The augmented version at least produces some detections (e.g., F1-score of 0.73% on *FLAME_train* and 16.5% on *FireMan_test*), implying a minor improvement in generalization due to augmentation. However, the overall performance remains extremely poor, and augmentation only offers marginal benefits.

Generalization Capacity with FireMan dataset with Data Augmentation

The generalization of Models 1 and 2, when trained on FireMan with data augmentation are demonstrated in TABLE XX and XXII, respectively. TABLE XXI shows the Confusion Matrix Components of Model 1, The ones from Model 2 are in Appendix G.

TABLE XX: Model 1 - FireMan Generalization with Data Augmentation

Metric	Model 1: FireMan (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,2175	1,1675	1,2883	0,9781	1,5843	0,2175
Accuracy	94,40	63,54	59,61	69,69	50,00	94,40
Precision	94,40	63,54	59,61	69,69	50,00	94,40
Recall	100,00	100,00	100,00	100,00	100,00	100,00
F1–Score	97,12	77,70	74,70	82,14	66,67	97,12
AUC	53,51	50,31	49,92	49,64	49,62	53,51
FDR	100,00	100,00	100,00	100,00	100,00	100,00
EWR	5,60	36,46	40,39	30,31	50,00	5,60

TABLE XXI: Model 1 - FireMan Generalization with Data Augmentation (CM)

Metric	Model 1: FireMan (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
TP	556	25018	5137	1145	7780	556
TN	0	0	0	0	0	0
FP	33	14357	3480	498	7780	33
FN	0	0	0	0	0	0

At first glance, Model 1 appears to have strong results, with perfect Recall (100%) across all datasets and reasonably high F1-scores. However, a closer inspection reveals a critical flaw: the model never classified any image as "no-fire." In every case, TN and FN are zero, and all predictions are positive ("fire"). As a result, Recall is artificially inflated, and the model fails to perform genuine classification. For example, in *FLAME_train*, the model produced 25018 true positives and 14357 false positives, resulting in an Accuracy of 63,54%, a Precision of 63,54%, and an EWR of 36,46%. Despite the seemingly strong F1-score of 77,70%, the model is simply predicting "fire" for every image. This pattern continues in *FLAME_test*, *UAVS_RawImages*, and *UAVS_AugImages*. When looking at the results in the "Split Test" (which in this test are the same as "*FireMan_test*"), the model performed poorly. Although it achieved 100% Recall, once again, this was due to the fact that it labelled every image as "fire", failing to identify any no-fire instances. This indicates that the model did not learn to distinguish between the two classes, even within its native domain, and highlights a critical flaw in its training and evaluation process.

TABLE XXII: Model 2 - FireMan Generalization with Data Augmentation

Metric	Model 2: FireMan (with Data Augmentation)					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,1438	3,4069	1,5613	1,1113	1,7886	0,1438
Accuracy	95,93	19,92	56,45	71,33	52,75	95,93
Precision	100,00	27,5	59,89	70,85	51,41	100,00
Recall	95,68	15,91	81,57	100,00	100,00	95,68
F1–Score	97,79	20,16	69,09	82,94	67,91	97,79
AUC	96,83	21,43	53,39	41,97	59,47	96,83
FDR	95,68	15,91	81,57	100,00	100,00	95,68
EWR	4,07	80,08	43,55	28,67	47,25	4,07

On *UAVS_RawImages*, the model performed strongly, achieving a perfect Recall of 100%, a F1-Score of 82,94%, and an Accuracy of 71,33%, demonstrating its capacity to detect fires in unfamiliar environments reliably. However, performance declined on *UAVS_AugImages*, with Accuracy dropping to 52,75% and EWR increasing to 47,25%. Performance on the FLAME datasets was notably weaker. On *FLAME_train*, the model achieved an Accuracy of 19,92%, an F1-Score of 20,16%, and a very high EWR of 80,08%. Even

on *FLAME_test*, metrics remained modest (F1-Score: 69,09%, EWR: 43,55%).

The non-augmented model (VI-B) consistently outperformed the augmented version on *FLAME_train* (Accuracy: 54,19% vs. 19,92%, F1-Score: 69,44% vs 20,16%) and *FLAME_test* (F1-Score: 72,87% vs 69,09%). Performance on UAVS datasets was comparable, though the augmented model achieved 100% Recall with slightly higher EWR.

Results Overview

The results from Phase 3 indicate that the impact of data augmentation on generalization differs notably between models. Model 2, which integrates attention mechanisms, showed substantial improvements in Recall, F1-score, and Accuracy across nearly all test scenarios, confirming its increased robustness to unseen data when trained with augmented samples. In contrast, Model 1 exhibited only marginal gains and continued to struggle, particularly with the FireMan dataset.

These findings suggest that data augmentation is most effective when used in combination with architectures that incorporate attention mechanisms. As such, for practical wildfire detection applications, data augmentation should be prioritized, particularly when deploying more complex, attention-based models. For simpler architectures, its benefits may be limited unless paired with additional strategies to improve feature extraction and generalization.

D. Phase 4: Dataset Combination Strategy

The final phase of the experimental framework examines how merging multiple datasets impacts the generalization capacity of UAV-based wildfire detection models. A model's resilience in unseen environments can be significantly increased by exposing it to more visual diversity during training. To explore this, Models 1 and 2 will be trained on four dataset combinations: FLAME and UAVs-FFDB, FLAME and FireMan, UAVs-FFDB and FireMan, and FLAME with UAVs-FFDB and FireMan, as explained in Subsection V-C.

Generalization Capacity with FLAME and UAVS-FFDB

The generalization of Models 1 and 2, when trained with FLAME and UAVS-FFDB are demonstrated in TABLE XXIII and XXIV, respectively. The confusion matrix components for each model are presented in Appendix H.

TABLE XXIII: Model 1 - FLAME and UAVS-FFDB

Metric	Model 1: FLAME and UAVS-FFDB					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,1121	0,0571	0,2993	0,0083	0,1083	6,5792
Accuracy	96,45	98,25	89,16	99,57	96,62	13,41
Precision	96,64	98,25	89,33	99,39	97,39	100,00
Recall	96,26	99,00	92,91	100,00	95,81	8,27
F1-Score	96,45	98,63	91,09	99,70	96,59	15,28
AUC	99,36	99,77	96,15	100,00	99,31	16,44
FDR	96,26	99,00	92,91	100,00	95,81	8,27
EWR	3,55	1,75	10,84	0,43	3,38	86,59

Model 1 performed well on semi-seen data, achieving high Accuracy on *FLAME_train* (98,25%) and strong results on *FLAME_test* (89,16% Accuracy, 92,91% Recall). *UAVS_RawImages* showed near-perfect performance (99,57% Accuracy, 100% Recall), indicating effective learning within familiar data domains. In contrast, performance on the entirely unseen *FireMan_test* dataset revealed poor generalization. The model achieved an Accuracy of just 13,41%, with a Recall of only 8,27% and an extremely high EWR of 86,59%. Despite a perfect Precision of 100%, the model failed to detect the vast majority of fire cases, highlighting its limited ability to adapt to new, out-of-domain data distributions.

TABLE XXIV: Model 2 - FLAME and UAVS-FFDB

Metric	Model 2: FLAME and UAVS-FFDB					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0543	0,0428	0,2196	0,0015	0,0468	7,1404
Accuracy	98,19	98,92	93,93	100,00	98,25	5,6
Precision	99,59	99,54	95,86	100,00	100,0	0,00
Recall	96,78	98,76	93,87	100,00	96,5	0,00
F1-Score	98,17	99,15	94,86	100,00	98,22	0,00
AUC	99,88	99,89	97,98	100,00	99,99	0,73
FDR	96,78	98,76	93,87	100,00	96,5	0,00
EWR	1,81	1,08	6,07	0,00	1,75	94,4

With 98,92% Accuracy and 98,76% FDR on *FLAME_train*, 93,93% Accuracy on *FLAME_test* and perfect scores on all metrics of *UAVS_RawImages* dataset, Model 2 demonstrates to operate well on semi-seen data. On the other hand, the model is unable to generalize to completely unknown data. It achieved a mere 5,60% Accuracy, 0% FDR, and 94,40% EWR on the *FireMan_test* dataset. Its inability to identify any fire occurrences (TP = 0, FN = 556) revealed a significant flaw in its capacity to manage unusual or varied input conditions.

Generalization Capacity with FLAME and FireMan

TABLE XXV and XXVI show how Model 1 and 2 generalize when trained using FLAME and FireMan dataset, respectively. Appendix I contains the confusion matrix components for each model.

TABLE XXV: Model 1 - FLAME and FireMan

Metric	Model 1: FLAME and FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0214	0,0179	0,0845	1,2169	1,8314	0,0124
Accuracy	99,38	99,45	97,67	71,64	60,87	99,66
Precision	99,47	99,32	97,85	73,43	56,97	100,00
Recall	99,73	99,81	98,25	92,93	88,83	99,64
F1-Score	99,60	99,57	98,05	82,04	69,42	99,82
AUC	99,93	99,98	99,51	64,47	57,97	100,00
FDR	99,73	99,81	98,25	92,93	88,83	99,64
EWR	0,62	0,55	2,33	28,36	39,13	0,34

Compared to models trained solely on FLAME or FireMan, the combined Model 1 showed significantly improved generalization to the unseen UAVS datasets. On *UAVS_RawImages*, it achieved 71,64% Accuracy, 92,93% Recall, and reduced the EWR to 28,36%, outperforming the FLAME-only model (22,82% Accuracy, 7,25% Recall) and FireMan-only model (62,99% Accuracy, 88,82% Recall). Similarly, on *UAVS_AugImages*, the combined model improved Accuracy (60,87%) and Recall (88,63%), with a lower EWR (39,13%) compared to the single-dataset models.

TABLE XXVI: Model 2 - FLAME and FireMan

Metric	Model 2: FLAME and FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0196	0,0156	0,0713	2,8197	3,3616	0,0067
Accuracy	99,73	99,49	97,54	30,98	20,25	100,00
Precision	99,82	99,44	98,73	50,68	18,89	100,00
Recall	99,82	99,77	97,12	35,90	18,06	100,00
F1-Score	99,82	99,60	97,92	42,02	18,46	100,00
AUC	99,90	99,97	99,76	17,93	11,38	100,00
FDR	99,82	99,77	97,12	35,9	18,06	100,00
EWR	0,27	0,51	2,46	69,02	79,75	0,00

The FLAME and FireMan datasets did not improve Model 2's generalization to the UAVS datasets. In comparison to the model trained exclusively on FireMan, which achieved 72,31% Accuracy, 89,08% Recall, and a significantly lower EWR of 27,69%, Model 2 only managed 30,98% Accuracy, 42,02% F1-score, and an EWR of 69,02% on *UAVS_RawImages*. On *UAVS_AugImages*, a similar trend is observed, with the FireMan-only model (59,06% Accuracy, 79,95% Recall) outperforming the mixed model (20,25% Accuracy, 18,46% Recall).

Generalization Capacity with UAVS-FFDB and FireMan

The generalization of Models 1 and 2 when trained using the UAVS-FFDB and FireMan dataset is demonstrated in TABLE XXVII and XXVIII, respectively. The confusion matrix components for each model are included in Appendiz J.

TABLE XXVII: Model 1 - UAVS-FFDB and FireMan

Metric	Model 1: UAVS-FFDB and FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,1176	1,9976	3,4718	0,069	0,1291	0,1791
Accuracy	96,42	68,85	56,71	98,11	96,4	93,04
Precision	95,71	67,41	58,57	97,36	94,31	98,68
Recall	98,47	98,69	93,56	100,00	98,77	93,88
F1-Score	97,07	80,11	72,04	98,66	96,48	96,22
AUC	99,36	79,72	42,85	99,97	99,46	95,05
FDR	98,47	98,69	93,56	100,00	98,77	93,88
EWR	3,58	31,15	43,29	1,89	3,06	6,96

When comparing the models, the one trained solely on the FireMan dataset demonstrates the best generalization to the unseen FLAME data. On *FLAME_train*, it achieves 79,50% Accuracy, 75,68% Precision, and 99,82% Recall, outperforming the combined UAVS and FireMan Model 1 with 68,85% Accuracy, 67,41% Precision and 98,69% Recall. A similar trend is observed on *FLAME_test*, where the FireMan-only model achieves 60,64% Accuracy and 99,38% Recall, compared to 56,71% Accuracy and 93,56% Recall in the combined model. While merging the datasets significantly improves over the UAVS-only model, which performs poorly on FLAME, it does not surpass the FireMan-only model, which remains the strongest in terms of generalization to this domain. Nevertheless, the combined model still achieves strong performance, indicating its robustness across different data sources.

TABLE XXVIII: Model 2 - UAVS-FFDB and FireMan

Metric	Model 2: UAVS-FFDB and FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0185	6,7210	3,7542	0,0043	0,0182	0,0459
Accuracy	99,65	13,42	33,94	100,00	99,68	98,98
Precision	99,76	21,29	40,39	100,00	99,81	100,00
Recall	99,65	13,44	22,72	100,00	99,55	98,92
F1-Score	99,71	16,48	29,08	100,00	99,68	99,46
AUC	99,94	8,40	33,62	100,00	99,93	99,64
FDR	99,65	13,44	22,72	100,00	99,55	98,92
EWR	0,35	86,58	66,06	0,00	0,32	1,02

Although combining UAVS and FireMan improved performance compared to using UAVS alone, it did not outperform the FireMan-only model on the FLAME dataset. On *FLAME_train*, the combined Model 2 achieved only 13,42% Accuracy, 13,44% Recall, and an F1-score of 16,48%, far below the FireMan-only model, which reached 54,19% Accuracy and 69,44% F1-score. On the *FLAME_test*, the combined model improved slightly (33,94% Accuracy, 22,72% Recall, F1-score 29,08%) but again remained below the FireMan-only model (61,77% Accuracy, 86,10% Recall, F1-score 72,87%). Model 2's overall generalization is poor on both *FLAME_train* and *FLAME_test*, which is significantly lower than on in-domain data, indicating limited robustness to unseen environments.

Generalization Capacity with FLAME + UAVS-FFDB + FireMan dataset

TABLE XXIX and TABLE XXX show how Models 1 and 2 generalize when trained with the FLAME, UAVS-FFDB and FireMan datasets. Appendiz K contains the confusion matrix components for each model.

TABLE XXIX: Model 1 - FLAME + UAVS-FFDB + FireMan

Metric	Model 1: FLAME + UAVS-FFDB + FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,1261	0,06300	0,2778	0,0032	0,1275	0,6246
Accuracy	95,50	97,72	90,45	99,94	95,29	82,17
Precision	95,46	97,44	90,03	99,91	93,73	100,00
Recall	96,93	99,01	94,43	100,00	97,07	81,12
F1-Score	96,19	98,22	92,18	99,96	95,37	89,57
AUC	99,21	99,77	96,06	100,0	99,29	91,70
FDR	96,93	99,01	94,43	100,00	97,07	81,12
EWR	4,50	2,28	9,55	0,06	4,71	17,83

Model 1 demonstrates strong generalization capabilities when applied to semi-seen datasets. On *FLAME_train*, it achieved 97,72% Accuracy, 97,44% Precision, and 99,01% Recall, with an EWR of just 2,28%. Although performance declined slightly on the *FLAME_test* (90,45% Accuracy, 94,43% Recall, EWR 9,55%), it remained solid. Results on *UAVS_RawImages* were nearly perfect, with 99,94% Accuracy, 100% Recall, and an EWR of only 0,06%. The model also performed reliably on *UAVS_AugImages*, achieving 95,29% Accuracy, 97,07% Recall, and an EWR of 4,71%.

TABLE XXX: Model 2 - FLAME + UAVS-FFDB + FireMan

Metric	Model 2: FLAME + UAVS-FFDB + FireMan					
	Split Test	FLAME_train	FLAME_test	UAVS_RawImages	UAVS_AugImages	FireMan_test
Loss	0,0348	0,0691	0,2187	0,0018	0,0228	0,0156
Accuracy	99,31	98,30	92,54	100,00	99,72	100,00
Precision	99,65	99,51	95,62	100,00	99,97	100,00
Recall	99,17	97,80	91,69	100,00	99,46	100,00
F1-Score	99,41	98,65	93,61	100,00	99,72	100,00
AUC	99,85	99,90	97,49	100,00	99,93	100,00
FDR	99,17	97,80	91,69	100,00	99,46	100,00
EWR	0,69	1,70	7,46	0,00	0,28	0,00

Model 2 demonstrates strong classification performance across semi-seen datasets. For example, *UAVS_RawImages* and *UAVS_AugImages* achieved nearly perfect results, with Accuracy, Precision, Recall, and F1-score all reaching or approaching 100% and EWR values of 0,0% and 0,28%, respectively. Similarly, *FLAME_train* demonstrated high performance (Accuracy: 98,30%, Recall: 97,80%, EWR: 1,70%), whereas *FLAME_test*, although derived from the same dataset but with different types of images, yielded a lower Accuracy of 92,54% and a notably higher EWR of 7,46%, indicating relatively weaker generalization. On *FireMan_test*, a split not used during training but from a dataset partially seen by the model, performance remained exceptional, with 100% across all metrics and 0.0% EWR, suggesting strong generalization to similar but previously unobserved data.

Results Overview

The result of phase 4 validates the intuitive expectation that the most robust and generalizable model is produced by training with all three datasets: FLAME, UAVs-FFDB, and FireMan. This configuration consistently performed better than all others, especially for Model 2, which showed strong adaptability to heterogeneous environments and achieved nearly perfect metrics across both seen and unseen test sets (*FireMan_test*).

The FLAME and FireMan dataset combination was the most successful of the other three, particularly for Model 1, which performed well when applied to UAV-based data. The FLAME and UAVs-FFDB combination, on the other hand, performed the worst, with neither model being able to generalize well to the FireMan dataset. This highlights the limitations that arise when training data lacks sufficient contextual and spatial variability.

A closer comparison between Model 1 and 2 reveals important differences. When trained on a variety of datasets, Model 2, which integrates attention mechanisms (mCBAM), performed better. However, in more limited training setups with less diverse datasets, it struggled, and occasionally failed entirely to detect the positive class. In contrast, Model 1 showed higher baseline consistency and frequently produced more balanced results in combinations of partial datasets. This suggests that while Model 2 performs best when given richly diverse training data and leverages its architectural complexity, Model 1 is more resilient in low-variance scenarios.

VII. Conclusion

The thorough investigation and rigorous experimentation described in this work make it clear that AI, and CNNs in particular, hold a great deal of potential for improving UAV image-based wildfire detection. Through systemic evaluation across multiple datasets, the report exhaustively investigates the ability of two state-of-the-art models to generalize beyond their training conditions. The findings show that although baseline replication works well, generalization to new data remains very difficult, particularly when training is performed on small or visually uniform datasets. When exploring generalization strategies, two primary approaches were tested: data augmentation and dataset combination.

Data augmentation yielded divergent results between Models 1 and 2. Being a model that incorporates attention mechanisms, Model 2 demonstrated significant improvements in generalization performance, highlighting its ability to leverage the increased variability introduced by augmentation. However, despite slight gains on some datasets, Model 1 frequently showed instability, with augmentation occasionally impairing performance. This was especially evident when the model was trained using the FireMan dataset, where overfitting to the positive class made it impossible to distinguish between fires and no-fire events.

When exposed to a variety of data, both models demonstrated improvements in terms of the dataset combination strategy despite producing completely different responses. Only when trained using the complete FLAME, UAVS-FFDB, and FireMan combination did Model 2 produce better results, achieving nearly flawless scores on every test set. However, on every other combination, especially when trained on less diverse data, Model 2 struggled to generalize and often missed the positive class. Contrarily, Model 1 demonstrated better generalization, even on partial combinations, and exhibited more consistent and well-rounded performance overall. While Model 2 works best in rich, heterogeneous training conditions, Model 1 showed higher resilience in constrained or homogeneous scenarios, suggesting a trade-off between robustness and complexity.

Regarding dataset quality for generalization, FireMan stood out as the most effective training set overall. Models trained on FireMan consistently demonstrated better adaptability across other datasets. On the contrary, UAVS-FFDB demonstrated the least generalization strength. Its limited environmental diversity and less complex scenes may have limited the models' learning capacity, as models trained exclusively on it did not perform well on FLAME or FireMan. In conclusion, UAVS- FFDB is the least valuable dataset for cross-domain generalization when used alone, whereas FireMan is the most valuable dataset.

These findings highlight the importance of dataset diversity and architecture selection in developing AI-based wildfire detection systems suitable for real-world applications. Although complex models are capable of achieving remarkable results, their efficacy is primarily reliant on how representative and variable the training data is. However, in certain situations, simpler architectures may offer more reliable performance. Future research should focus on creating standardized, highly annotated UAV datasets and investigating models that are both lightweight and flexible to ensure high accuracy and generalizability in operational settings.

REFERENCES

[1] K. Rao, A. P. Williams, J. F. Flefil, and A. G. Konings, "Sar-enhanced mapping of live fuel moisture content," *Remote Sensing of Environment*, vol. 245, p. 111797, 2020.

[2] S. Giannakidou, P. Radoglou-Grammatikis, T. Lagkas, V. Argyriou, S. Goudos, E. K. Markakis, and P. Sarigiannidis, "Leveraging the power of internet of things and artificial intelligence in forest fire prevention, detection, and restoration: A comprehensive survey," *Internet of Things*, vol. 26, p. 101171, 2024.

[3] S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, P. Fule, A. Watts, N.-M. T. Kokolakis, and K. G. Vamvoudakis, "A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," *Information Fusion*, vol. 108, p. 102369, 2024.

[4] M. N. Ramadan, T. Basmaji, A. Gad, H. Hamdan, B. T. Akgün, M. A. Ali, M. Alkhedher, and M. Ghazal, "Towards early forest fire detection and prevention using ai-powered drones and the iot," *Internet of Things*, vol. 27, p. 101248, 2024.

[5] M. Costa, "Project's github repository," https://github.com/Maguids/RuralFireDetection.

[6] M. N. Mowla, D. Asadi, S. Masum, and K. Rabie, "Adaptive hierarchical multi-headed convolutional neural network with modified convolutional block attention for aerial forest fire detection," *IEEE Access*, vol. 13, pp. 3412–3433, 2025.

[7] S. Chaturvedi, C. S. Arun, P. S. Thakur, P. Khanna, and A. Ojha, "Ultra-lightweight convolution-transformer network for early fire smoke detection," *Fire Ecology*, vol. 20, p. 83, 2024.

[8] S. Rajalakshmi, Sellam, N. Kannan, and S. Saranya, "Exploiting drone images for forest fire detection using metaheuristics with deep learning model," *Global NEST Journal*, vol. 25, no. 7, pp. 147–154, 2023.

[9] B. C. Arrue, A. Ollero, and J. R. Martinez de Dios, "An intelligent system for false alarm reduction in infrared forest-fire detection," *IEEE Intelligent Systems and Their Applications*, vol. 15, no. 3, pp. 64–73, 2000.

[10] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Fulé, and E. Blasch, "The flame dataset: Aerial imagery pile burn detection using drones (uavs)," 2020. [Online]. Available: https://dx.doi.org/10.21227/qad6-r683

[11] M. N. Mowla and D. Asadi, "Uavs-fddb: Uavs-based forest fire detection database," 2024. [Online]. Available: https://doi.org/10.17632/5m98kvdkyt.3

[12] A. Shamsosoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: The flame dataset," *Computer Networks*, vol. 193, p. 108001, 2021.

[13] M. N. Mowla, D. Asadi, K. N. Tekeoglu, S. Masum, and K. Rabie, "Uavs-ffdb: A high-resolution dataset for advancing forest fire detection and monitoring using unmanned aerial vehicles (uavs)," *Data in Brief*, vol. 55, p. 110706, 2024.

[14] P. Skalski, "Makesense.ai," https://www.makesense.ai/.

[15] W. Kularatne, K. Sedaghat Shayegan, C. Álvarez Casado, J. Rajala, T. Hänninen, M. Bordallo López, and N. L. Nguyen, "Fireman-uav-rgbt (1.0.0-beta.2) [data set]," 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13732947

APPENDIX A
BASELINE REPRODUCTION - FLAME: CONFUSION MATRICES

TABLE I: Confusion Matrix - Model 1

Actual Class	**Predicted: Positive**	**Predicted: Negative**
Positive	284	1
Negative	5	271

TABLE II: Confusion Matrix - Model 2

Actual Class	**Predicted: Positive**	**Predicted: Negative**
Positive	7688	92
Negative	0	7780

APPENDIX B

PHASE 2: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FLAME DATASET

TABLE III: Model 1: Confusion Matrix Components of FLAME

Metric	**Model 1: Confusion Matrix Components of FLAME**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	284	24895	5081	83	1700	60
TN	271	14317	3378	292	4420	33
FP	5	40	102	206	3360	0
FN	1	123	56	1062	6080	496

TABLE IV: Model 2: Confusion Matrix Components of FLAME

Metric	**Model 2: Confusion Matrix Components of FLAME**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	283	24922	5040	261	1512	50
TN	275	14320	3450	404	6472	33
FP	1	37	30	94	1308	0
FN	2	96	97	884	6268	506

APPENDIX C
PHASE 2: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF UAVS-FFDB DATASET

TABLE V: Model 1: Confusion Matrix Components of UAVS-FFDB

Metric	**Model 1: Confusion Matrix Components of UAVS-FFDB**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	7595	101	356	1140	7595	328
TN	7469	7245	3325	497	7469	0
FP	311	7112	155	1	311	33
FN	185	24917	4781	5	185	228

TABLE VI: Model 2: Confusion Matrix Components of UAVS-FFDB

Metric	**Model 2: Confusion Matrix Components of UAVS-FFDB**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	7688	27	0	1145	7688	0
TN	7780	14230	3471	498	7780	33
FP	0	127	9	0	0	0
FN	92	24991	5137	0	92	556

APPENDIX D
PHASE 2: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FIREMAN DATASET

TABLE VII: Model 1: Confusion Matrix Components of FireMan

Metric	**Model 1: Confusion Matrix Components of FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	556	24974	5105	1017	5617	556
TN	33	6331	120	18	240	33
FP	0	8026	3360	480	7540	0
FN	0	44	32	128	2163	0

TABLE VIII: Model 2: Confusion Matrix Components of FireMan

Metric	**Model 2: Confusion Matrix Components of FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	509	20494	4423	1020	6220	509
TN	33	845	900	168	2969	33
FP	0	13512	2580	330	4811	0
FN	47	4524	714	125	1560	47

APPENDIX E

PHASE 3: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FLAME DATASET WITH DATA AUGMENTATION

TABLE IX: Model 1: Confusion Matrix Components of FLAME with Data Augmentation

Metric	**Model 1: Confusion Matrix Components of FLAME with Data Augmentation**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	280	24755	4704	564	2265	21
TN	247	12930	2435	409	5542	33
FP	29	1427	1045	89	2238	0
FN	6	263	433	581	5515	535

TABLE X: Model 2 - Confusion Matrix Components of FLAME with Data Augmentation

Metric	**Model 2: Confusion Matrix Components of FLAME with Data Augmentation**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	278	24370	4791	612	4868	494
TN	271	14040	2922	320	4378	0
FP	5	317	558	178	3402	33
FN	8	648	346	533	2912	63

APPENDIX F
PHASE 3: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF UAVS-FFDB DATASET WITH DATA AUGMENTATION

TABLE XI: Model 1 - Confusion Matrix Components of UAVS-FFDB with Data Augmentation

Metric	**Model 1: Confusion Matrix Components of UAVS-FFDB with Data Augmentation**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
Loss	0,0242	4,5953	5,8676	0,0333	0,0242	7,4245
TP	7776	3313	1383	1141	7776	97
TN	7700	8157	1592	490	7700	0
FP	80	6200	1888	8	80	33
FN	4	21705	3754	4	4	459

TABLE XII: Model 2 - Confusion Matrix Components of UAVS-FFDB with Data Augmentation

Metric	**Model 2: Confusion Matrix Components of UAVS-FFDB with Data Augmentation**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
Loss	0,0088	9,0587	8,2179	0,0108	0,0088	7,261
TP	7780	98	0	1145	7780	50
TN	7764	12605	3468	495	7764	33
FP	16	1752	12	3	16	0
FN	0	24920	5137	0	0	506

APPENDIX G

PHASE 3: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FIREMAN DATASET WITH DATA AUGMENTATION

TABLE XIII: Model 2 - Confusion Matrix Components of FireMan with Data Augmentation

Metric	**Model 2: Confusion Matrix Components of FireMan with Data Augmentation**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	532	3980	4190	1145	7780	532
TN	33	3864	674	27	428	33
FP	0	10493	2806	471	7352	0
FN	24	21038	974	0	0	24

APPENDIX H
PHASE 4: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FLAME + UAVS

TABLE XIV: Model 1: Confusion Matrix Components of FLAME + UAVS-FFDB

Metric	**Model 1: Confusion Matrix Components of FLAME + UAVS-FFDB**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	9418	24768	4773	1145	7454	46
TN	9421	13917	2910	491	7580	33
FP	327	440	570	7	200	0
FN	366	250	364	0	326	510

TABLE XV: Model 2: Confusion Matrix Components of FLAME + UAVS-FFDB

Metric	**Model 2: Confusion Matrix Components of FLAME + UAVS-FFDB**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	9469	24709	4822	1145	7508	0
TN	9709	14242	3272	498	7780	33
FP	39	115	208	0	0	0
FN	315	309	315	0	272	556

APPENDIX I
PHASE 4: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FLAME + FIREMAN

TABLE XVI: Model 1: Confusion Matrix Components of FLAME + FireMan

Metric	**Model 1: Confusion Matrix Components of FLAME + FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	1120	24970	5047	1064	6911	554
TN	331	14187	3369	113	2561	33
FP	6	170	111	385	5219	0
FN	3	48	90	81	869	2

TABLE XVII: Model 2: Confusion Matrix Components of FLAME + FireMan

Metric	**Model 2: Confusion Matrix Components of FLAME + FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	1121	24960	4989	411	1405	556
TN	335	14216	3416	98	1746	33
FP	2	141	64	400	6034	0
FN	2	58	148	734	6375	0

APPENDIX J
PHASE 4: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF UAVS + FIREMAN

TABLE XVIII: Model 1: Confusion Matrix Components of UAVS-FFDB + FireMan

Metric	**Model 1: Confusion Matrix Components of UAVS-FFDB + FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	12103	24691	4806	1145	7684	522
TN	7596	2420	81	467	7316	26
FP	543	11937	3399	31	464	7
FN	188	327	331	0	96	34

TABLE XIX: Model 2: Confusion Matrix Components of UAVS-FFDB + FireMan

Metric	**Model 2: Confusion Matrix Components of UAVS-FFDB + FireMan**					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	12248	3363	1167	1145	7745	550
TN	8110	1921	1758	498	7765	33
FP	29	12436	1722	0	15	0
FN	43	21655	3970	0	35	6

APPENDIX K
PHASE 4: CONFUSION MATRIX COMPONENTS OF GENERALIZATION CAPACITY OF FLAME + UAVS + FIREMAN

TABLE XX: Model 1: Confusion Matrix Components of FLAME + UAVS-FFDB + FireMan

Metric	Model 1: Confusion Matrix Components of FLAME + UAVS-FFDB + FireMan					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	13856	24770	4851	1145	7552	451
TN	9448	13706	2943	497	7275	33
FP	659	651	537	1	505	0
FN	439	248	286	0	228	105

TABLE XXI: Model 2: Confusion Matrix Components of FLAME + UAVS-FFDB + FireMan

Metric	Model 2: Confusion Matrix Components of FLAME + UAVS-FFDB + FireMan					
	Split Test	**FLAME_train**	**FLAME_test**	**UAVS_RawImages**	**UAVS_AugImages**	**FireMan_test**
TP	14177	24468	4710	1145	7738	556
TN	10057	14237	3264	498	7778	33
FP	50	120	216	0	2	0
FN	118	550	427	0	42	0