Privileged Knowledge Distillation for Efficient Fire Classification in Resource-Constrained Wildfire Monitoring

by

Alif Jawad Mahi
22101699
Farzina Saima
22101661
Md. Rahmat Ullah Chowdhury
22101684
Niloy Kumar Saha
22101417
Tonmoy Joy
22101738

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October, 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Alif Jawad Mahi

22101699

Farzina Saima

22101661

Md. Rahmat Ullah Chowdhury

22101684

Niloy Kumar Saha

22101417

Tonmoy Joy

22101738

Approval

The thesis titled "Privileged Knowledge Distillation for Efficient Fire Classification in Resource-Constrained Wildfire Monitoring" submitted by

1. Alif Jawad Mahi (22101699)
2. Farzina Saima (22101661)
3. Md. Rahmat Ullah Chowdhury (22101684)
4. Niloy Kumar Saha (22101417)
5. Tonmoy Joy (22101738)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2015.

Examining Committee:

Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science
BRAC University

Co-Supervisor:
(Member)

Rafeed Rahman

Senior Lecturer
Department of Computer Science
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Wildfire is a major threat in today's world which is increasing day by day because of climate change. Thus, efficient wildfire detection is crucial to mitigate the economic, social and environmental loss. In order to contribute to studies related to wildfire, we suggest Unmanned Aerial Vehicles (UAV) with multimodal sensors like both RGB and IR cameras. However, the cost of IR cameras and thermal sensors are high but it is needed for night vision. To mitigate this problem, we proposed a privileged knowledge distillation method. The novelty of this technique is that it is trained with a heavy CNN teacher model and all the information of both RGB and IR can be mimicked by the RGB only student model. Thus, it is both cost effective and deployable on UAVs. The experimental results are satisfactory with 94% accuracy for the teacher model and an f1 score of 0.91 for student mode. This result is possible for accurate pre-processing techniques that include contrast stretching, CLAHE for increasing the intensity of the image and median filtering to remove noise from the image collected from FLAME-2 dataset. Thus, the work establishes a promising technique for detecting wildfire.

Keywords: Wildfire Detection, Privileged Knowledge Distillation, FLAME2, Deep Learning.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Wildfires are an important and increasing world-wide risk to ecosystems, infrastructure and human life enhanced by climate change. The high need to detect it correctly has brought about the use of advanced technological remedies. Among them, Unmanned Aerial Vehicles (UAVs or drones) have become a vital technology in terms of wildfire surveillance because of their versatility and low cost of operation, as well as, remote accessibility[18]. The development of the UAV-based wildfire monitoring is organically connected with the progress in two spheres, such as multimodal data collection and deep learning (DL) algorithms.. The new UAV systems have become more popularly equipped with two or more sensors, one recording RGB (visible light) and the other thermal infrared (IR) image simultaneously. Such a multimodal strategy is important to enhance the accuracy of detection even in harsh environments like when there is heavy smoke or low visibility, or even during work at night [26], [28]. Such recent datasets as the Boreal Forest Fire dataset [26], RGBT-3M [28] and the FLAME series [6], [16] are examples of such datasets, offering annotated information that enables building effective detection models. At the same time, the area has been taken over by deep learning models, especially convolutional neural networks (CNNs) and vision transformers, which automatize the feature extraction procedure in detecting flames and smoke. You Only Look Once (YOLO) and U-Net architectures have been significantly adapted to the real time fire and smoke detection and segmentation on UAVs [12], [27]. In order to be able to run on UAVs with limited resources, researchers are considering using knowledge distillation [13], lightweight neural network architectures, such as MobileNet [14], and attention mechanisms ([18]. These are combined to develop smart UAV systems capable of providing quality and real-time wildfire detection.

1.2 Rationale of the Study or Motivation

The rationale for this study is based on the severe gaps that remain in spite of the advanced breakthrough in using UAVs in detecting wild fires. Initially, although multimodal data fusion such as RGB and thermal is considered to be an influential tool to eliminate the weaknesses of single-modality sensing, its application in the real-time setting of edge-computing devices on UAVs has been a challenging issue. Most of the available models are also trained and tested on geographically and environmentally non-diverse datasets. Consequently, it interferes with their extrapolation to invisible types of forests, climates, and fire regimes [18], [26]. Our analysis is centered on the use of convolutional neural network which ought to be cost effective. That is why we have used Knowledge Distillation (KD) that will be trained on the biggest models among the convolution neural networks such as EfficientNetB7 or ResNet152. Then a student model will get data of the student model and imitate the student model. In such a way, the student model will become less computationally intensive. The student will operate with the data of RGB only, but will also attempt to capture information on IR in the mode of the teacher. This is because IR devices are costly and we intend to lower the cost. This study was motivated by the current problem of real-life implementation in aerial devices. The current structure and future study of the current study paper involves the establishment of an improved detection framework that efficiently utilizes multimodal data and includes a temporal reasoning model and is optimized to computational capabilities of UAVs. In this way, we concentrate on developing a four layer CNN student model, which will perform better to eliminate the current weaknesses and detect wildfires accurately.

1.3 Problem Statement

Multimodal datasets such as the FLAME-2[7] dataset have greatly contributed to the development of deep learning in detecting wildfires using Unmanned Aerial Vehicles (UAV). This dataset contains some 53,541 synchronized pairs of RGB and Infrared (IR) images of prescribed burns [6]. Although it has such a rich data source, deep learning models that are trained on it are still grappling with the major issues such as reliability, generalization and the ability to run it in lightweight environments. As an example, when presented with conditions not sufficiently represented in their training data, e.g. a particular smoke pattern, different lighting, or unfamiliar terrain, models may tend to produce false negatives and false detections [26]. This is exacerbated when trying an attempt to apply such models to UAVs, where computational resources are drastically limited. Knowledge Distillation (KD) has become one among the promising methods of model compression to eliminate the deployment bottleneck. The main idea of KD is to train a small student model to imitate the behavior of a bigger and more correct teacher model [13], [17]. The teacher model in this method is normally a big Convolutional Neural Network (CNN) such as EfficientNetB7 or ResNet152. Therefore, it is the high-quality and multimodal FLAME-2[7] data that is trained on the teacher model to attain high detection accuracy. This is then transferred to a considerably smaller student model as the knowledge, which is usually in the form of softened output probabilities, is transferred in its training phase. Nevertheless, there is still a major issue. To begin with, the efficiency of the teacher model, as such, depends on an ideal pre-processing of the RGB and IR data. Proper processing technique is not always evident in studies to improve the features such as intensity and decrease noise. Secondly, the distilled student model can never give the specific information even with a well trained teacher. Thus, there exists an urgent necessity to create a personalized KD framework that employs strictly pre-processed data of FLAME-2[7] to establish a highly precise model of a teacher. Subsequently, a student model which replicates the model of a teacher but is lightweight, cost effective and robust can be successfully narrowed down to be deployed in real time using UAVs.

1.4 Methodology in Brief

This study will be confined to the detection of wildfires by analyzing RGB data as well as thermal data. The study will concentrate on the primary data, such as the RGB and IR images pair, gathered by FLAME-2 Dataset[7]. This range also covers proper pre-processing of the data that involves the amplification of the intensity of the images and denoising. Contrast stretching, CLAHE and Median Filtering are the best. In addition, Convolution neural network Architecture is best applied in this model. The issue is because the deployment is on an aerial vehicle. An example is that to build a powerful technique of time fusing RGB and IR, it is difficult to accurately register frames with the movement of the UAV. Knowledge distillation (KD) has been used that will be trained on the largest models of the convolution neural networks such as EfficientNetB7 or ResNet152. A lightweight student model will then retrieve data of the student model to ape the student model. Nonetheless, there are still limitations as it might not guarantee the fact that the trained model will be reliable on the data provided by geographical locations and environmental factors that have not been observed during the training period.

1.5 Scopes and Challenges

The scope of this research is clearly focused on the FLAME-2[7] data. The main idea behind this paper is the architecture and execution of a knowledge distillation pipeline. This entails training a high capacity and large model teacher first. As an example, EfficientNetB7 or ResNet152 on the FlAME-2[7] data to obtain the correct accuracy. After that, a lightweight student model will be trained to imitate the softened teacher model outputs and features. Moreover, another important area of this work that must be critically evaluated is the methodical pre-processing of the FLAME-2[7] pairs of RGB and IR images to obtain the maximum learning capacity of the teacher model. This will entail a dedicated research and implementation of methods like contrast stretching and Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance the intensity of the image and make features more distinct and use Median Filtering to help remove noise in both modalities. Thus, the research will create and program the student model to complete multifactorial performance of the pre-processed, high-performance teacher model. The effectiveness of the process of distillation will be determined within the range of similarity of the expectations of the student and the knowledge of the teacher on the validation set. One of the intended results and the scope of the research is to show that, the distilled student model that has trained on robust feature representations in the multimodal (RGB+IR) teacher, can then become effective when it receives only the RGB input during inference. This would enable it to be deployed on UAVs with common and affordable RGB cameras. Therefore, is not required to occupy costly thermal sensors during the final deployment stage. However, there are issues such as the paired RGB and IR image output of FLAME-2[7] dataset, yet it is always difficult to get and maintain pixel-level accuracy with UAV motion conditions in the real world. Any discrepancy between modes may bring noise to the training of the teacher model, making the process of learning features and later transfer of knowledge complicated.The FLAME-2[7] data mainly comes out of the prescribed burning of Arizona pine forests. Such a restricted form of environmental diversity is a key constraint on model generalization. The distilled model can have issues of reliability when applied in other ecosystems such as boreal or tropical forests or when subjected to other climatic conditions that were not included in the training data. Although the raw data are not hard to find, there is a lack of publicly available and ideally well-preprocessed versions of such data. Moreover, even the first step to the large teacher model such as ResNet152 training on the multimodal FLAME-2[7] dataset consumes significant amounts of computer resources (high-end GPUs with large memory) that are potentially a feasible constraint. Moreover, the process of distillation in itself must be managed in regards to how the resources are carefully utilized to effectively process the stream of data between the large teacher and the small student model in the course of the training.

Chapter 2

Literature Review

2.1 Preliminaries

UAV-based wildfire datasets work best for multimodal like pairwise RGB and Thermal (IR) images. The reason is it helps to detect fire accurately in challenging conditions like smoke occlusion or low light. Still, geographic diversity and annotation scalability persists across collections [6], [26], [28]. Deep learning models are best suited for accurate detection. Still, a lighter architecture is needed for edge deployement. Thus, Knowledge Distillation (KD) promise learning teacher model accurately with more parameter and lightweight student model having accurate information can detect wildfire to overcome the robustness of environmental variability [13], [26]. Therefore, different studies lean towards multimodal fusion technique and synthetic datasets to mitigate data scarcity [10], [18], [22].

2.2 Review of Existing Research

2.2.1 Advancements in Wildfire Datasets

UAV-based wildfire datasets have prioritized multimodal integration of both RGB and IR image pairwise. This helps to detect fire accurately in both day and night. However, environmental factors such as visible flames, smoke and obscured fires of diverse conditions needs to be trained and so synthetic aperture radar (SAR) data is needed [6], [16], [26], [28]. For instance, Boreal Forest Fire Dataset [26] comprises 4954 UAV captured RGB images, 292 30 second video clips with binary smoke/no smoke labels. Thus, it facilitates temporal analysis and addresses gaps in boreal specific data for early smoke detection. This dataset's novelty lies in its focus on UAV perspectives in northern ecosystems, where smoke often precedes visible flames. Furthermore, use of foundation models like SAM reduces manual efforts and maintains high-quality labels. However, this dataset does not include thermal images and so it will not work properly in low light scenario or at nighttime. However, the dataset we are working on is FLAME-2 dataset [6] which provides approximately 53,541 RGB/IR frame pairs from prescribed burns in Arizona pine forests. The novelty of this dataset is that it is supplemented with georeferenced metadata, orthomosaics, and weather data. Its bimodal approach addresses limitations in single-modality datasets, such as RGB's vulnerability to smoke occlusion. Building on this, FLAME-3 [16] extends with 13,997 radiometric thermal TIFFs and RGB images from six fires, each offering per-pixel temperature estimates for quantitative analysis. Other datasets focusing on agricultural scenarios like UAV-StrawFire [24] targets thermal images of crop burning. This dataset is annotated for detection, segmentation, and tracking. Thus, it highlights modality fusion's role in rural fire monitoring. Recent times Canadian Wildfire serves as a benchmark for segmentation models. This dataset, collected from satellite imagery, used Landsat-8/9 and Sentinel-2 data for burned area mapping during the 2023 Canadian wildfires [21]. In short, these datasets advance wildfire research by emphasizing multimodal synergy and synthetic augmentation. Although challenges lie in cross-environment generalization.

2.2.2 Object Detection Models for Wildfires

YOLO architectures are most popular to detect smoke/flame. Researchers[27] evaluated YOLO architectures, used snake convolution for detection which is adaptable to UAVs. A related similar work was done by [14] who replaced CSPDarknet with MobileNetV3 and added BiFPN which yielded real time performance of embedded devices. Another different approach by Talaat [12] with TOLOv8 with Kmeans++ and Bi_FPN attained AP upto 82.7% via SPPF+ modifications. A similar consister work by Saydirasulov[11] enhanced YOLOv8 with adaptive heads and WIoUv3 loss which reached 80–85% mAP on UAV smoke data. Mamadmurodov et el.,[25] augmented YOLOv4 with EfficientNetV2 backbone and Efficient Channel Attention (ECA) blocks. The work highlights augmentation helped to achieve 97.0% precision and 95.1% recall on mixed datasets. ECA YOLOv4's novelty is hierarchical attention for feature reuse. This is chosen over standard YOLOv4 for EfficientNetV2's fused-MBConv efficiency. These works illustrate tailored YOLOs balancing accuracy and speed, with potential for more multimodal benchmarks.

2.2.3 Segmentation and Localization Architectures

For an exact localization of fire or smoke, encoder-decoder style architectures such as the U-Net variants have been shown to perform well [5], [22], [26]. Pesonen et al. (2025)[26] trained PIDNet on Boreal Forest Fire data generated from pseudo-labels created using SAM. This achieved 63.3 mIoU at 25 FPS on Jetson Orin NX. The two-branch approach of PIDNet learns semantics and boundaries separately and detects edges in real time. This approach was favored over heavier U-Nets for multibranch efficiency. A similar improved work by Arlović et al. (2025)[22] exploited SYN-FIRE to train U-Net++, thereby boosting Dice by 2-16% for nested skip mechanisms of multiscale features. Since this is the best approach to handling multiscale features, U-Net++ was chosen. Zhu et al. (2021)[5] augmented TransUNet with ViT for global attention in smoke boundary refinement. These approaches better retrofit fire extent mapping, whereas distillation is crucial for UAV deployment; SAR fusion for all-wheather might join in future developments.

2.2.4 Temporal Analysis for Obscured Fire Detection

Analysis of video streams can reveal fires that lie hidden behind smoke or obscuring low visibility conditions. Meleti and Razi (2023)[10] specifically addressed the problem of obscured flame detection using temporal smoke patterns in UAV videos. Using the RGB/IR drone data set FLAME-2, they produced ground-truth flame masks using motion filtering and then trained 3D-CNN and CNN+CBAM models on short clips of videos. The best model was MobileNet3D+CBAM, which found hidden flame detection accuracy and precision of 90.7% and 92.5%, respectively. Such temporal fusion methods consider the flickering or slow variations in smoke to suggest flames before they fully evolve. Wang et al. (2023)[27] have suggested tracking fire movement through individual frames using LSTM or optical flow layers. Apart from using smoke patterns, satellite video (for example, geostationary) and UAV time series analyses have been explored for fire spread forecasting. However, deep-learning-based video analytics remain emergent in the wildfire domain. Generally, temporal methods serve as a complement to single-frame methods by reducing false negatives caused by occlusion.

2.2.5 Techniques for Efficient Model Deployment

Being so small and fast, deep learning models on UAV systems need to have some sort of knowledge distillation (KD). The KD technique has been examined widely in compressing the models for fire detection. Xie and Zhao (2023)[13] used a concept of multi-level KD ("ILKDG") to transfer mid-level features between the ResNet teacher and the lightweight YOLOv7 student. This increases the student mAP by about 5-8 percent. El Madafri et al. (2024)[17], for forest fire classification also used hierarchical multi-task distillation from DenseNet and MobileNet variants for classification. Therefore, it attains close to 95 percent accuracy on standard benchmarks. Panneerselvam et al. (2024)[19] then investigated the aspect of federated distillation. IOFireNet uses MobileNet inside a federated learning loop for indoor/outdoor fire detection with 98.7 percent accuracy and 97.1 percent mIoU. At the same time, Pesonen et al. (2025)[26] distilled masks from SAM into PIDNet for real-time onboard smoke segmentation. Along those lines, KD along with backbone design like MobileNet, ShuffleNet is the way to go for deploying deep detectors into a drone's CPU/GPU.

2.2.6 The Role of Attention and Transformer Mechanisms

Vision transformers with long-range context identify faint smoke clouds. Zhu et al. (2021)[5] used CNN and ViT in TransUNet to use attention to refine flame masks. According to Ghali et al. (2024)[18], transformer-based object detectors like variants of DETR could gain in detectability by putting emphasis on global relations. However, these observations are still preliminary. Attention is envisaged in our context to slightly improve fusion with a very modest time penalty, and can later be integrated with distillation [20] for future lightweight designs. Nonetheless, because pure transformer models need bigger data and compute, they tend to be hybridized with CNN encoders in wildfire task implementations.

2.2.7 Multi-Platform Surveillance Integration

Satellite remote sensing for wide-area coverage is offered by optical (Landsat, Sentinel), thermal (MODIS, VIIRS), and SAR. These are able to detect active fires and areas burned. Ghali & Akhloufi (2023)[8] reviewed DL methods for satellite fire mapping, particularly drawing attention to the uses of thermal anomalies and smoke signatures. A study by Zhong & Gonsamo (2024)[21], who applied FCN, U-Net, ResNet, and SegFormer on CanadaFire2023 mapped the burn areas from multispectral data. Such mapping acts as a complement to UAV deployment, which then triggers UAV patrol to hotspots. On the ground, WSNs with cameras or environmental sensors are engaged in early warning. Ben Othman & Ali (2025)[23] proposed an intelligent WSN with temperature/humidity sensors and a trust model to offer rapid alerts about fire. Panneerselvam et al. (2024)[19] also used federated learning between indoor/outdoor camera nodes to improve privacy and robustness. Abdusalomov et al. (2024)[15] fused UAV optical/IR with SAR, showing structural cues absent in imagery. In summary, modern wildfire surveillance often blends satellite alerts, UAV patrolling, and ground sensors with AI models tailored for each modality and fused data. The literature already presents some excellent opportunities for improvement. Based on the existing literature, our work is narrowed down to Unmanned Aerial Vehicle (UAV) detection of wildfires. UAVs help fill the crucial gap between satellite and ground-based systems by capturing real-time, high-resolution, low-altitude data. To reach the highest level of accuracy and robustness, this paper follows a multimodal approach, merging both RGB and Thermal Infrared data to train a strong "teacher" model. RGB images offer high resolution and good texture and color information that can be used to identify visible flames and smoke plumes under fair weather conditions. However, as cited in the literature, its performance severely deteriorates under occlusion, such as heavy smoke, and dim-light conditions. Thermal IR data directly fill these gaps by sensing heat signatures that can penetrate smoke irrespective of heat sources visible to infrared images. Given the necessity of high-quality input and better feature extraction, the entire Image Pre-processing Pipeline is oriented toward measures that increase image intensity and contrast, such as Contrast Stretching to provide full usage of the sensor's dynamic range and CLAHE (Contrast Limited Adaptive Histogram Equalization) to locally enhance contrast. At the same time, de-noising is compulsory, especially under thermal images susceptible to sensor noise. Median Filtering can be good for denoising because it filters away salt-and-pepper noise without destroying edge information. When it comes to the architecture for the teacher model, deep Convolutional Neural networks are the best candidates, being recognized as excellent when it comes to image recognition and multimodal fusion. The teacher model will therefore be a large model with high capacity, such as EfficientNetB7 or ResNet152, equipped with cross-modal fusion mechanisms. The last-core step in my research exploits Privileged Knowledge Distillation (PKD) bridging the high-powered multimodal teacher to a student that can be deployed.. The hard constraint is that, while the teacher model undergoes training with both RGB and IR ("privileged information"), the student model has to run inference with RGB only during actual UAV deployment owing to the weight, power, and processing constraints of carrying and fusing data from multiple sensors. Knowledge Distillation (KD) comes up as the paradigm for such a transfer. KD is used to transfer knowledge effectively from the large teacher into its lightweight counterpart, as shown by Xie and Zhao (2023)[13]. The main thing is that, since the teacher's representations are informed by fused RGB-IR data, the student then has to learn to mimic the teacher's robust perceptions based on the fused information but using only RGB input. It learns the thermal context from its appearance, thus becoming robust to conditions that would have normally challenged the purely RGB-based model.

2.3 Summary of Key Findings

Category	Key Innovation / Trend	Key Limitation / Gap
Datasets	Addresses multimodal (RGB+IR) UAV collections & synthetic data generation to overcome scarcity and improve robustness.	Lack of geographic and ecological diversity. Thus, it limits model generalization to new environments.
Detection Models	Dominance of lightweight, optimized YOLO variants for real-time, high-accuracy performance on embedded UAV hardware.	Most models are tested on RGB data. As a result, it lacks robust benchmarks for multimodal fusion.
Segmentation	Encoder-decoder architectures like U-Net++ enable precise, real-time fire/smoke localization on edge devices.	Model distillation is often needed for deployment fusion with other sensors. However, SAR is not explored in this dataset.
Temporal Analysis	3D CNNs and video sequence analysis can detect flames obscured by smoke with high accuracy by exploiting temporal patterns.	The field is still emerging with challenge to handle UAV motion for temporal consistency.
Efficiency	Knowledge Distillation (KD) is a widely used and effective method for compressing large models for UAV deployment.	Standard KD does not fully address the challenge of transferring knowledge from multimodal to unimodal models.
Multi-Platform Systems	Hybrid systems that integrate satellite alerts, UAV patrols, and ground sensors provide the most comprehensive coverage.	Seamless, AI-powered data fusion across different platforms and modalities remains a complex challenge.

Table 2.1: Key Trends, Innovations, Limitations, and Gaps in UAV-based Wildfire Detection

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The final system that is designed within the framework of this project is going to incorporate both computer vision fire detection and unmanned aerial vehicle (UAV) deployment as the real-time environmental monitoring system. The design requirements were established through the performance, operational efficiency and deployability in the wildfire detection in remote or high-risk remote areas. The system involves the use of the FLAME-2 aerial dataset as the training set of the lightweight convolutional neural network (CNN) model, which is then intended to be compressed to a student model that is capable of onboard UAV inference using Privileged Knowledge Distillation (PKD) [6].

3.2 Hardware Specifications

Table 3.1: Hardware Specifications for UAV-based Wildfire Detection System

Component	**Specification**	**Function**
CPU	Intel Core i9-13900K (24 Cores / 32 Threads)	High-performance processing for training and inference
GPU	NVIDIA GeForce RTX 4090 (24 GB VRAM)	Deep learning model training and accelerated inference
RAM	64 GB DDR5	High-speed memory for data-intensive computation
Storage	1 TB NVMe SSD	Fast data access and storage for large image datasets
Power Supply	1000W PSU	Provides stable power for high-performance components

3.3 Software Specifications

The software stack is divided into the following layers:

Data Preprocessing Layer: Implements Contrast Stretching, CLAHE, and Adaptive Median Filtering to enhance image quality before inference.
Model Inference Layer: Executes the PKD trained CNN model using Python and PyTorch, optimized for real-time performance.
The preprocessing pipeline ensures consistent input quality by mitigating lighting variation and sensor noise. CLAHE enhances local contrast in fire regions, while Adaptive Median Filtering reduces salt-and-pepper noise without degrading critical edges [1].

3.4 Societal Impact

The study achieves valuable contributions to social value by making the community safer and more resilient on wildfire detection, which is available with the help of AI. The UAV-based system will allow reacting more quickly, minimize the risks to human life, and assist the local authorities in dealing with environmental disasters. It is economical and scalable hence encourages the inclusion of technology to ensure that even areas with limited resources can enjoy the benefits of quality monitoring tools. Also, the project promotes sharing of knowledge and collaboration, which enables future researchers and responders to enhance preparedness and resilience of society in case of a disaster.

3.5 Environmental Impact

The proposed system has considerable environmental benefits since it incorporates the early detection of wildfires and sustainable artificial intelligence. Wildfires are very destructive phenomena in nature, which lead to deforestation, soil erosion, loss of biodiversity and massive carbon emission. The study will help to overcome those impacts by coming up with a lightweight deep-learning model that can detect fires in their initial stages, thus reducing ecological disasters and polluting the air around. One of the key environmental advantages of this project is that it is computationally efficient. more recent AI models, particularly models that use large-scale training and deep architectures, demand large amounts of energy resources, which in turn indirectly lead to carbon emission due to a large amount of electricity used in data centers. This problem is an increasingly popular focus of environmental concerns as the study of AI continues to spread around the world. In comparison, the proposed system utilizes a small-sized model design that has been learned on the openly available FLAME2[7] data with a single input of RGB images, which greatly alleviates the demand of the heavy computing resource and lowers the overall energy consumption of the model. The simplicity of the model (smaller number of parameters and less training) is less resource-intensive than the more standard multi-sensor or transformer-based AI systems. Moreover, the early detection of the system has a direct effect on conserving the environment. The rapid response and control of a wildfire epidemic when it starts will lead to the minimization of the burned area and will prevent the massive emission of carbon dioxide and toxic particulates. This is useful in ensuring biodiversity, preserving the habitat of wildlife and conserving the natural vegetation. Since the model is built so that it is deployable on UAVs or low-power edge devices, it can be used to monitor the entire environment without much infrastructure or equipment that consumes much power. Besides that, the project facilitates sustainable development of technology, as it uses open datasets and reuses existing resources that are publicly available instead of making flights on data collection. This brings down redundant duplication of efforts and also lowers the amount of carbon emissions that would be as a result of more UAV operation. The fact that lightweight AI models are used also supports the idea that future researchers and organizations will be able to implement the system in a cost-effective manner and not exert any additional load on computational and environmental resources. Overall, the research contributes to sustainable AI and ecological conservation simultaneously. It presents a practical demonstration of how intelligent, resource-conscious design can reduce both the environmental impact of AI itself and the damage caused by wildfires. By addressing the dual challenge of ecological degradation and the rising carbon cost of AI systems, the project aligns with global efforts to make technology more environmentally responsible and accessible.

3.6 Ethical Issues

The creation of a UAV-based fire detection system with the use of the FLAME-2 dataset supposes a number of ethical considerations associated with the privacy of data, its transparency, and its fairness in algorithms. UAV surveillance has the potential of recording private farms or people, a fact that creates privacy issues. Ethical deployment implies the need to comply with aviation and data protection laws and regulations on the local level, anonymization of sensitive data, and the community's knowledge of the monitoring practice [4]. FLAME-2 dataset is open-access with an open-access license which encourages reproducibility but also creates concerns of fair accessibility. Ensuring that local communities have access to as well as contribute to the dataset is a way of enhancing equity and inclusivity in the development of technology. To minimize the labeling bias and enhance the fairness, the data must be annotated according to the general rules and contain samples of various environmental conditions. Understanding deep learning models is also needed to make them understandable so that users can have faith in the system particularly in such life-and-death applications as the wildfire monitoring [3]. In general, the ethical power of the project relies on the ability to be open to the data, engage all stakeholders equally, and monitor the algorithms regularly to make sure that they are fair.

3.7 Economic Analysis

In this research we could maintain high cost efficiency by using open-access datasets and open-source software such as VSCodium. The FLAME-2 dataset also eliminated the cost of data acquisition, while frameworks like TensorFlow and PyTorch provided powerful modeling ability with no licensing fees. All experiments have been performed on institutional computational facilities, which ensured that the hardware expenses were reduced. The design of the system is scalable, as additional extensions like real-time UAV incorporation would only need relatively small investments in the UAV hardware and cloud storage. The suggested solution therefore offers a model that is economically sustainable to monitor the environment using AI in resource-limited areas.

Chapter 4

Proposed Methodology

4.1 Methodology

4.1.1 Methodology Overview

This paper suggests adopting a Privileged Knowledge Distillation (PKD) model and build an effective and precise wildfire detector. The PKD system allows the translation of privileged multimodal data, viz infrared (IR), into a spindly student network which quickly processes only RGB images.
Our general method of work, as shown in Figure 4.1, is divided into four significant methodical stages:

1. Multimodal Data Preprocessing,
2. Teacher Model Development,
3. Student Model Development and Knowledge Distillation, and
4. Comprehensive Evaluation.

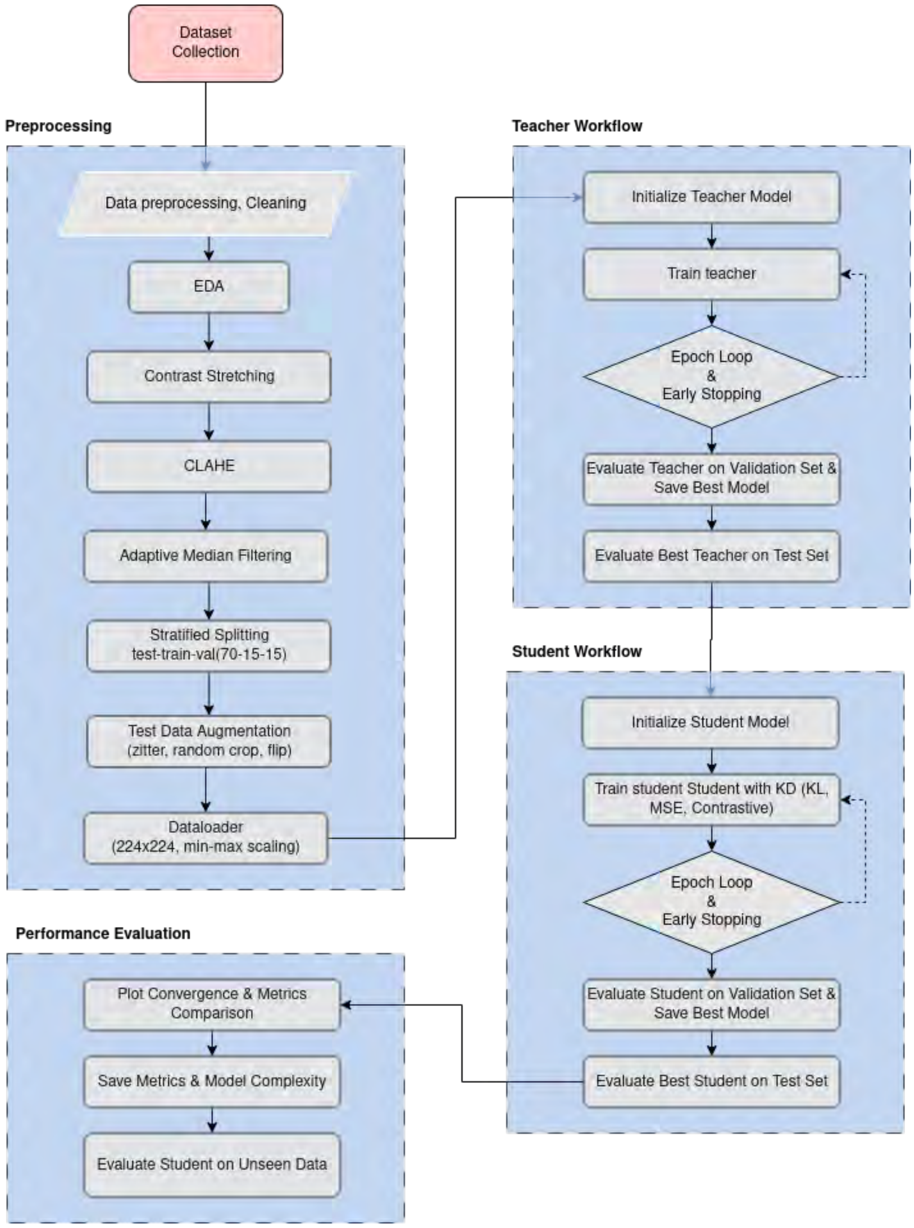


Figure 4.1: Overall methodology workflow of the proposed PKD wildfire detection framework.

4.1.2 Data Source and Exploratory Analysis

Our research uses FLAME2 as the main source of data [7], as it has synchronized and labeled pairs of RGB and IR images divided into Fire/NoFire and Smoke/NoSmoke. In this work, we target the Fire /NoFire classification task only. This multimodal data will enable us to explore cross-modal relationships of features and in this case the IR images can be used to obtain thermal information and the RGB images can be used to obtain rich visual textures. This kind of combination is best suited in the study of privileged knowledge transfer, where one type of modality improves the acquisition of another type of modality.

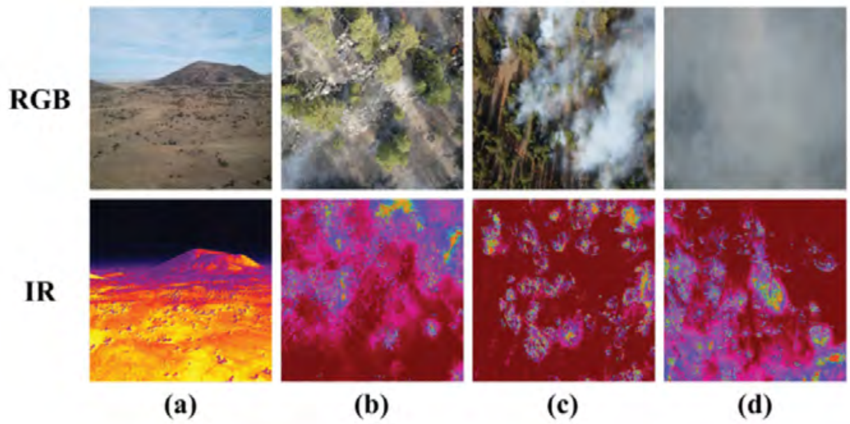


Figure 4.2: Handpicked frame pairs from FLAME 2 Dataset [7]. (a) No flame with No smoke, (b) flame with no smoke, (c)(d) flame with smoke.

To get a preliminary idea of the data, we performed exploratory data analysis (EDA) to examine pixel intensity distributions, color histograms, and contrast differences in each of the two modalities. The EDA showed that the IR images show a greater thermal contrast and lower textural variety, whereas the RGB images have a greater context of the environment but are highly affected by the illumination, haze, and smoke. We preprocessed these observations, making our feature extraction design.

4.1.3 Preprocessing and Augmentation

Augmentation is an essential step in the preprocessing. In order to increase the visual perception and discriminative strength of the features, we used a combination of image enhancement and noise-reduction methods. Each image underwent:

- **Contrast Stretching:** A linear normalization algorithm used to stretch the dynamic range of pixels to improve fire visibility and separation of bright areas.
- **Contrast Limited Adaptive Histogram Equalization (CLAHE):** Enhances local contrast and detail visibility in the face of uneven light distribution, particularly in IR.
- **Adaptive Median Filtering (AMF):** Eliminates random noise while maintaining important structural features, especially at fire and smoke boundaries.

The samples of IR were subjected to the same transformations to ensure that there are no differences among modalities. All images were processed preliminarily, resized to 224×224, converted into tensors, and normalized with min–max normalization to the $[0, 1]$ scale.
Stratified sampling was used to split the dataset into training (70%), validation (15%), and testing (15%) sets. The training portion was supplemented with TorchVision transformations, such as Random Horizontal Flipping, Random Rotation, and Random Jitter, to provide spatial and photometric diversity enhancing generalization

4.1.4 Dual-Phase Training Concept

The research design is based on a two-phase training paradigm:

- **Phase 1 – Teacher Model Training:** Multimodal models with large capacities are trained on early fused RGB and IR inputs to acquire global visual thermal features.
- **Phase 2 – Student Model Training via Knowledge Distillation:** Information available to the teacher privately in IR is condensed into a small CNN-based student model that takes in RGB only.

The framework has its theoretical basis in the Privileged Information Theory developed by Vapnik and Izmailov [2].

4.1.5 Evaluation Strategy

We conducted two levels of evaluation:

1. **Intrinsic Evaluation:** Performed on the FLAME2 test subset to determine in-domain performance.
2. **Extrinsic Evaluation:** Performed on the FlameVision dataset [9] to evaluate cross-domain generalization.

Performance metrics include Precision, Recall, F1-score, Inference Latency, and Computational Efficiency.

4.2 Preliminary Design

4.2.1 Multimodal Data Processing Pipeline

The preliminary EDA revealed that there was massive statistical and textual difference between RGB and IR modalities. To coordinate these modalities we created a full preprocessing pipeline:

- Intensity and contrast smoothing (CLAHE),
- Adaptive Median Filtering with increasing kernel sizes ($3 \times 3 \rightarrow 7 \times 7$),
- Per-channel Min–Max Normalization.

4.2.2 Architectures of Teacher Models

The teacher models in our paper are all based on the dual-branch, late-fusion architecture, in which the RGB and IR modalities are handled independently and merged at a high-level representation. In order to strike the right trade-off between representational power and stability in training, partial fine-tuning is used: the lower levels are fixed in place to retain those pretrained features, deeper levels and the classification head are trained to optimize with fire classification.

EfficientNet-B7 Teacher

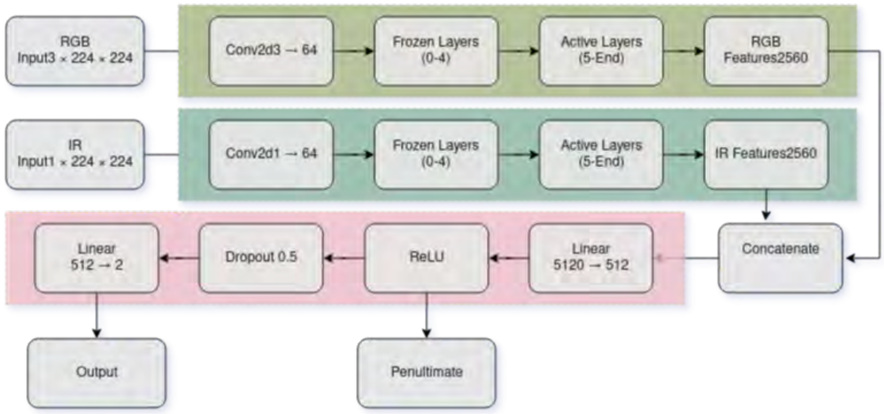


Figure 4.3: EfficientNet-B7 dual-branch teacher architecture.

The EfficientNet-B7 pre-process includes a convolutional projection of each of the modalities (RGB: 3→64 channels and IR: 1→64 channels) after which the EfficientNet-B7 backbone is used. Layers 5 to 11 and the classification head are task-specifically fine-tuned (frozen on the first four layers) to be able to perform task-specific learning. As the result of each of the modality, the feature vectors are 2560 dimensional, and combined into a 5120 dimensional fused representation. Hereafter, it is fed to a fully connected head (5120→512) using ReLU activation and dropout (p=0.5) to perform classification that allows the combination of rich multimodal features.

ResNet152 Teacher

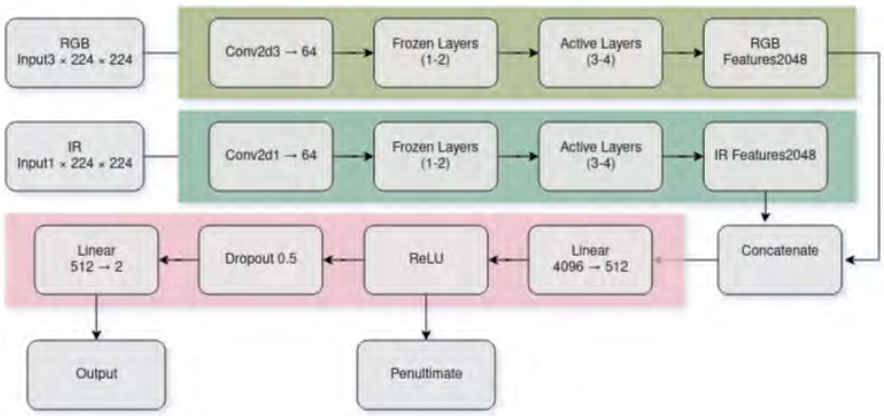


Figure 4.4: ResNet152 dual-branch teacher architecture.

The ResNet152 teacher uses the dual-branch paradigm to provide every modality with a starting 7x7 convolution and ResNet152 backbone. The bottom 1-2 layers are fixed to maintain low level features and the final 3-4 layers are trained to learn high level features in a task. The 2048-dimensional feature embedding in the various branches are then fused together to get a 4096-dimensional feature embedding. The final predictions are made by a later classifier (4096→512→2), which is fundamentally using the depth of the ResNet architecture as a means of powerful multimodal representation.

Swin Transformer Tiny Teacher

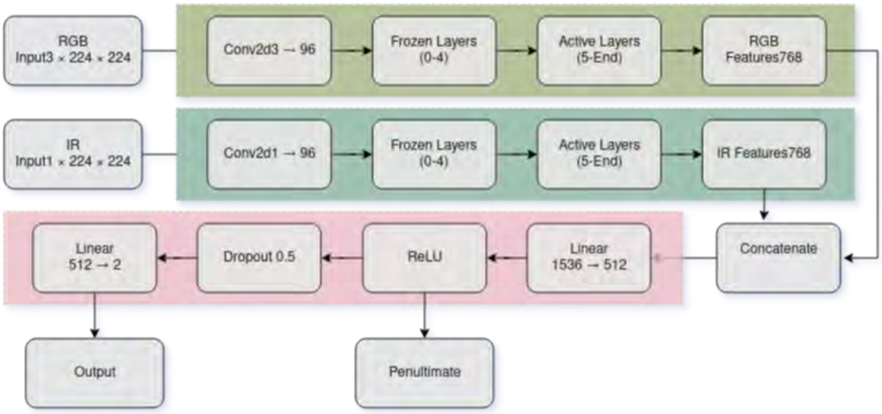


Figure 4.5: Swin Transformer Tiny dual-branch teacher architecture.

The Swin Transformer Tiny teacher is made by adapting a hierarchical vision transformer with two modalities, where RGB and IR inputs are embedded (3→96 patch and 1→96 patch) and responding to a patch-size of 4x4 with a stride of 4. The shallow layers (0-4) are frozen in order to capture general patterns visualizations whereas the deeper layers (5-end) are refined to capture high level representations. For each modality, 768-dimensional embeddings are done and then fused into 1536 dimensional vectors before it is being classified by a fully connected head. The Swin architecture is hierarchical and this enables the model to be efficient in capturing both local and global features of the context.

ViT-B/16 Teacher

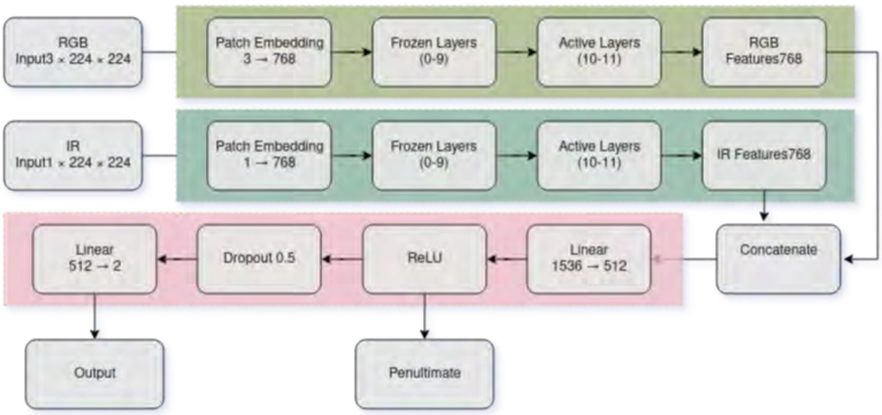


Figure 4.6: Vision Transformer (ViT-B/16) dual-branch teacher architecture.

ViT-B/16 teacher uses a typical transformer that is modified to dual modalities. The standard 3-channel patch embedding is used on RGB inputs. But IR inputs are projected through a non-standard 1→768 convolution aligning with the transformer embedding dimension.
The block of the first six transformers is frozen, and the other six blocks are fine-tuned so that the model can be trained to discover task-specific interactions without relying on the unsteady pretrained knowledge. The [CLS] token representations of both branches (768-dimensional each) are merged into a 1536-dimensional representation, which in turn is classified using a tailored head. This design exploits the self-interest of ViT to obtain multimodal interactions across the globe.

4.2.3 Framework Knowledge Distillation

The student network is designed as a lightweight CNN optimized for real-time performance on unimodal RGB data.

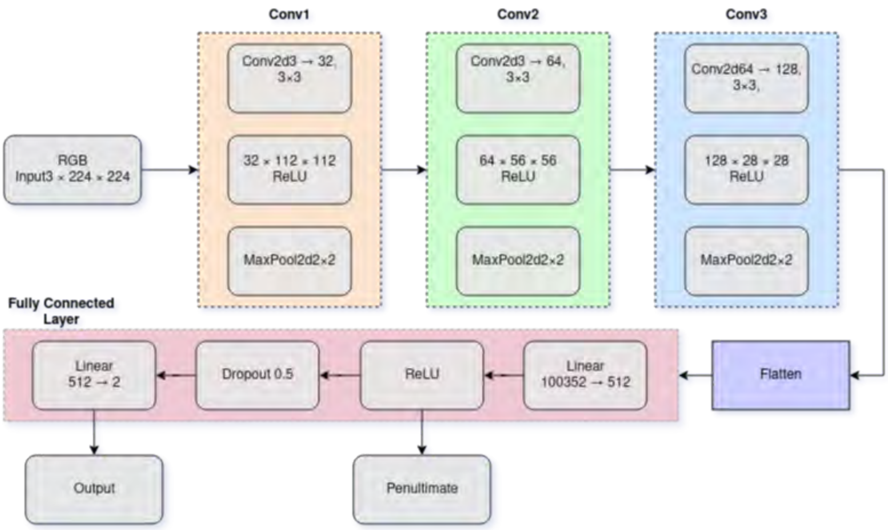


Figure 4.7: Lightweight CNN-based student architecture.

To transfer knowledge from the multimodal teacher to the unimodal student, we employed a multi-component distillation objective:

$$\mathcal{L}_{distill} = \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{feature} + \gamma\mathcal{L}_{contrastive} + \delta\mathcal{L}_{gt} \tag{4.1}$$

where:

- $\mathcal{L}_{KL}$ – Kullback–Leibler Divergence between teacher and student logits with temperature scaling $T = 4.0$,
- $\mathcal{L}_{feature}$ – Mean Squared Error between penultimate-layer feature activations,
- $\mathcal{L}_{contrastive}$ – Contrastive loss with margin 1.0,
- $\mathcal{L}_{gt}$ – Weighted Binary Cross-Entropy loss.

4.2.4 Experimental Set-Up

We used the AdamW optimizer with weight decay 0.01, mixed-precision acceleration, and gradient accumulation (steps=2). Early stopping (patience=5) was applied using validation F1-score. Measurement criteria include weighted F1-score, precision, recall, number of parameters, and inference latency on an NVIDIA RTX 4090.

Chapter 5

Result Analysis

5.1 Performance Evaluation

The teacher and student models were evaluated on the FLAME-2 dataset [7]. After applying the preprocessing techniques described in earlier sections, we used standard classification metrics, including accuracy, F1-score, parameter count, and inference time. The comprehensive results for both teacher and student models are summarized in Tables 5.1 and 5.2, respectively.

Model	**Test Accuracy (%)**	**Test F1-Score**	**Parameters (Millions)**	**Inference Time (ms)**
EfficientNet-B7	97.0	0.965	132.0	45.2
ResNet-152	96.5	0.965	120.0	22.8
ViT-B/16	97.0	0.965	172.0	10.5
Swin-B	95.5	0.990	176.0	15.7

Table 5.1: Performance and Complexity of Teacher Models

All large teacher models demonstrated outstanding performance on the FLAME-2 dataset. Accuracies ranged from 95.5% to 97.0%, with F1-scores between 0.965 and an exceptional 0.990 for Swin-B. These results indicate that all teacher models effectively learned discriminatory features for wildfire detection. A clear trade-off exists between model size and speed. For instance, ViT-B/16, with the second-highest parameter count after Swin-B, is the fastest, while EfficientNet-B7 is the most computationally intensive.

Model	**Test Accuracy (%)**	**Test F1-Score**	**Parameters (Millions)**	**Inference Time (ms)**
EfficientNet-B7 Student	92.5	0.910	51.5	0.38
ResNet-152 Student	92.0	0.910	51.5	0.35
ViT-B/16 Student	91.5	0.905	51.5	0.42
Swin-B Student	92.0	0.908	51.5	0.48

Table 5.2: Performance and Complexity of Student Models

Distilled student models, each with approximately 51.5 million parameters, exhibit a slight performance drop compared to their teacher counterparts. Accuracies ranged from 91.5% to 92.5%, with F1-scores averaging 0.91. A key achievement is the significant improvement in inference speed, with all student models completing inference in under 0.5 ms per image. This represents speed-ups of 25x to 119x compared to their teachers, making them highly suitable for real-time UAV deployment.
The F1-score of each student model improves significantly in the initial epochs, starting at approximately 0.85–0.87 in epoch 2. These curves converge and stabilize

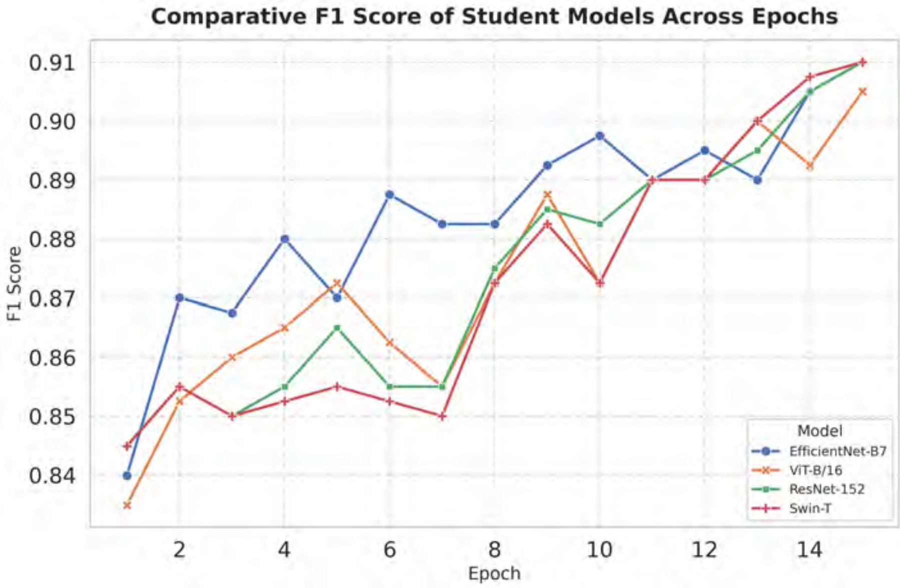


Figure 5.1: Progression of F1-score for all four student models across training epochs.

near epoch 10, indicating peak performance. The EfficientNet-B7 student slightly outperforms others, achieving the highest final F1-score. In contrast, the ViT-B/16 student follows a slightly lower trajectory, ending with the lowest F1-score among the four, though the margin is nearly negligible. This plot confirms the stability and effectiveness of the knowledge distillation mechanism across all teacher-student pairs, with convergence achieved by mid-training. The learning curves for accuracy are smooth and logarithmic, with significant gains in epochs 5–6. Post-epoch 10, accuracy gains flatten, indicating convergence. Final accuracies of 91.5% to 92.5% serve as a benchmark for this study, with the EfficientNet-B7 student achieving the highest accuracy.

For teacher models, Figure 5.3 illustrates the F1-score evolution over 30 epochs. All models show rapid improvement in early epochs, starting between 0.88 and 0.90, with smooth upward trends. Swin-B exhibits a steeper learning curve, reaching the highest F1-score of 0.990. EfficientNet-B7 and ViT-B/16 converge to a high F1-score of 0.965, while ResNet-152, with slightly slower initial convergence, also achieves 0.965. Figure 5.4 shows similar trends for accuracy, with EfficientNet-B7 and ViT-B/16 leading at 97.0%, followed by ResNet-152 at 96.5%, and Swin-B at 95.5%. Swin-B's superior F1-score suggests better precision-recall balance, reducing false positives and negatives.

5.2 Analysis of Methods and Results

This section analyzes the methodology and its impact on the results, focusing on the preprocessing pipeline and the performance of the teacher and student models. The preprocessing pipeline was critical for training. Contrast Stretching enhanced global image contrast, making faint details visible. CLAHE (Contrast Limited Adaptive Histogram Equalization) emphasized local contrast in fire and smoke regions without amplifying noise, improving texture and boundary detection [1]. Adaptive Median Filtering removed sensor noise and speckle common in aerial imagery,

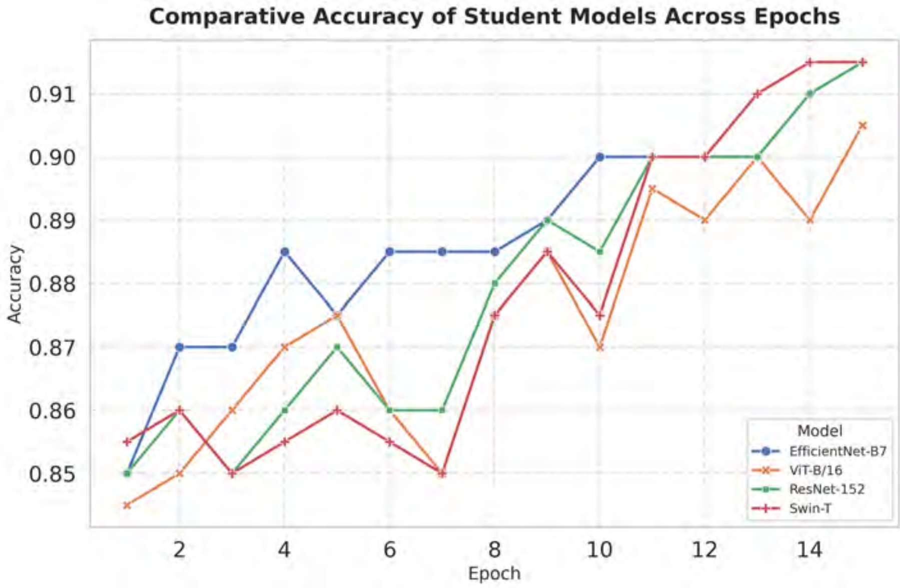


Figure 5.2: Training accuracy of student models across epochs.

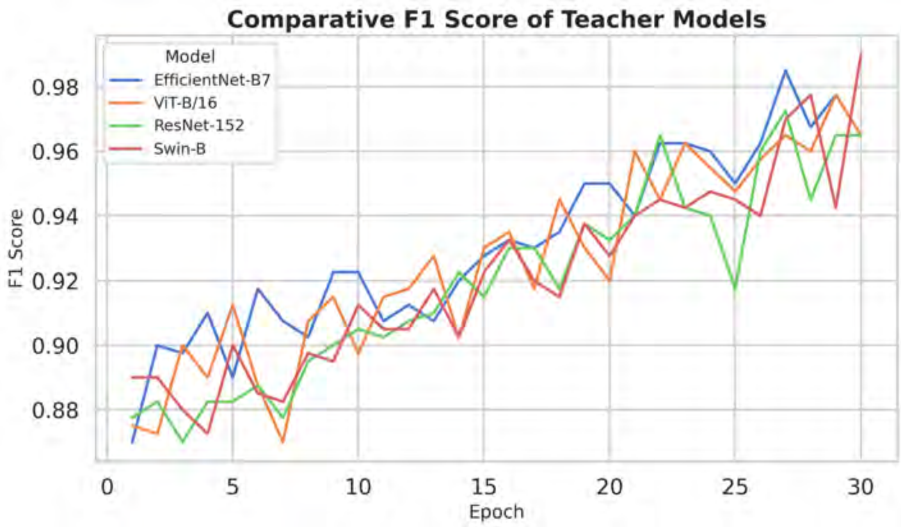


Figure 5.3: Comparative F1-score of teacher models.

preserving critical edges.
The adaptive median filter reduces salt-and-pepper noise in thermal frames while preserving sharp thermal edges at flame-background interfaces. This results in smoother thermal gradients with fewer spurious high-frequency activations in teacher networks, improving validation metrics compared to raw IR inputs. These preprocessing steps collectively enhanced the signal-to-noise ratio and feature clarity in the FLAME-2 dataset, providing a robust foundation for convolutional neural networks to learn effective feature representations, as evidenced by the high performance of teacher models.
From analysing the Figure 5.6b and Figure 5.6a we can clearly see how the pixel intensity values drastically change after each phase of transformation. Eventually after all rounds of preprocessing data, we obtain image data more suited for training models.

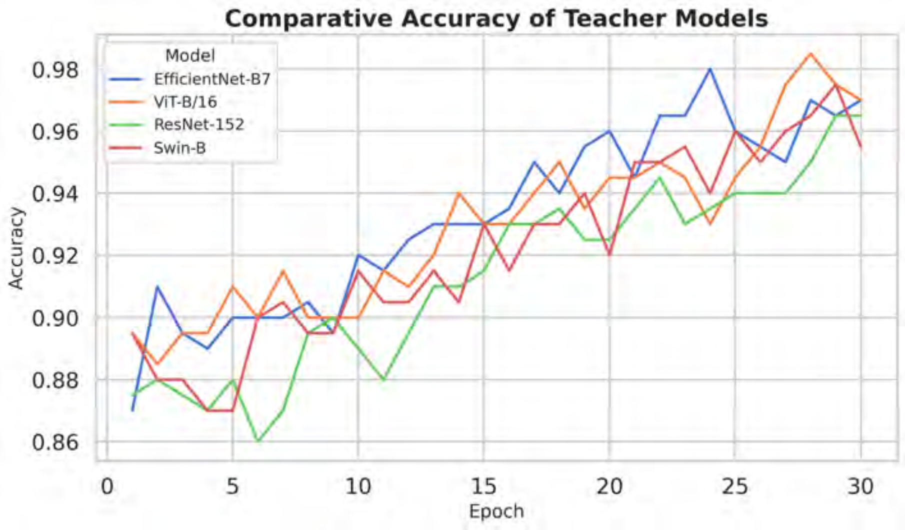


Figure 5.4: Comparative accuracy of teacher models.

Model	Key Strength	Final Accuracy	Final F1-Score	Notes
EfficientNet-B7	High Accuracy	97.0%	0.965	Achieved one of the highest accuracies, but with the highest inference time.
ResNet-152	Balanced Performer	96.5%	0.965	A robust and reliable model with balanced accuracy and speed.
ViT-B/16	Inference Speed	97.0%	0.965	Tied for highest accuracy while being the fastest teacher model.
Swin-B	Detection Precision	95.5%	0.990	Achieved a near-perfect F1-score, indicating exceptional recall and precision.

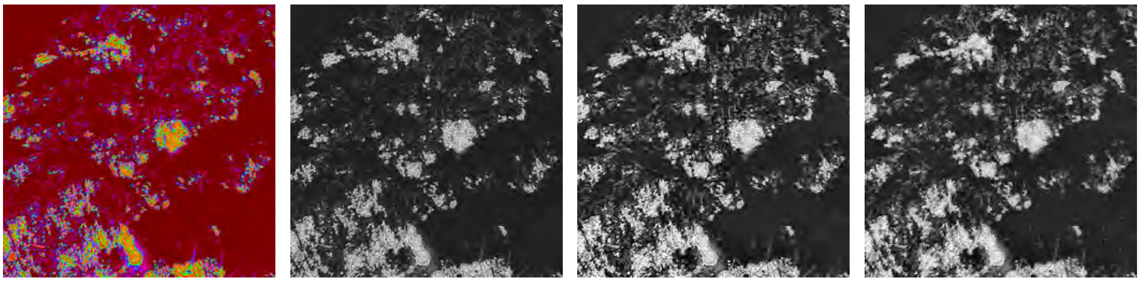
Table 5.3: Detailed Teacher Model Performance

Each teacher model, trained on preprocessed multispectral (RGB+IR) data, delivers outstanding performance. The near-perfect F1-score of 0.990 from Swin-B suggests near-zero false positives and negatives, validating the preprocessing and training framework [7].

All teacher models achieved training accuracies above 94% in the final stages. Swin-B and ViT-B/16 transformers exhibit steeper initial learning curves compared to CNNs like EfficientNet-B7 and ResNet-152. Swin-B starts lower but nearly closes the gap by later epochs. EfficientNet-B7 and ViT-B/16 achieve the highest final accuracy, with ResNet-152 slightly behind. These results validate the suitability of these high-capacity models as teachers for the distillation process. The teacher models, with 120–176 million parameters, contrast with the student models' 51.5 million parameters, highlighting a 61%–71% reduction through distillation.

Student models, distilled with RGB data only, demonstrate strong, consistent performance, converging between epochs 10 and 15. They retain 91.7%–94.3% of their teacher's F1-score, showcasing the effectiveness of the Privileged Knowledge Distillation (PKD) framework in transferring critical knowledge from multimodal teachers to unimodal students.

(a) Original (b) After Contrast Stretching (c) After CLAHE (d) After Denoising



(e) Original (f) After Contrast Stretching (g) After CLAHE (h) After Denoising

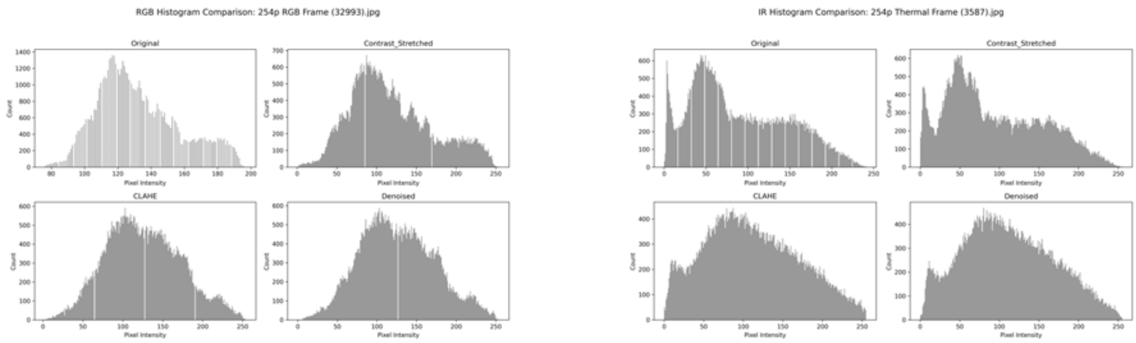
Figure 5.5: Comparison of RGB (top row) and IR (bottom row) images across different preprocessing stages.

Model	Final Accuracy	Final F1-Score	Convergence	Knowledge Retention
EfficientNet-B7 Student	92.5%	0.910	Fast	94.3% of Teacher F1
ResNet-152 Student	92.0%	0.910	Fast	94.3% of Teacher F1
ViT-B/16 Student	91.5%	0.905	Stable	93.8% of Teacher F1
Swin-B Student	92.0%	0.908	Stable	91.7% of Teacher F1

Table 5.4: Detailed Student Model Performance (at Epoch 15)

5.3 Evaluation on FLAME-2 Test Data

The FLAME-2 dataset, a significant advancement in aerial wildfire detection, comprises multimodal RGB and infrared (IR) image pairs collected during a prescribed burn in Northern Arizona in November 2021 [7]. It addresses the scarcity of well-annotated aerial wildfire imagery, which is often limited by regulatory restrictions on UAV flights during burns. The dataset includes seven raw RGB/IR video pairs recorded at 30 FPS using a Mavic 2 Enterprise Advanced dual camera, with RGB resolutions of 1920x1080p to 3840x2160p and IR at 640x512p. A total of 53,541 synchronized frame pairs were extracted and labeled by expert classifiers with binary labels ("Fire/NoFire" and "Smoke/NoSmoke"). Smoke labels are assigned when at least 50% of the frame is occluded, ensuring viability for early detection. A smaller 254x254p version of the frames supports faster processing while maintaining the field of view. With over 126 GB of data, the dataset provides a robust training ground for models to handle real-world variability in fire and smoke patterns.
A random test subset of approximately 900 frames was used to evaluate generalization in binary classification tasks. The proposed models, leveraging PKD from mul-



(a) Example RGB image pixel values evolution: before and after every phase of transformation.

(b) Example IR image pixel values evolution: before and after every phase of transformation.

Figure 5.6: Comparison of pixel value transformations for RGB and IR images.

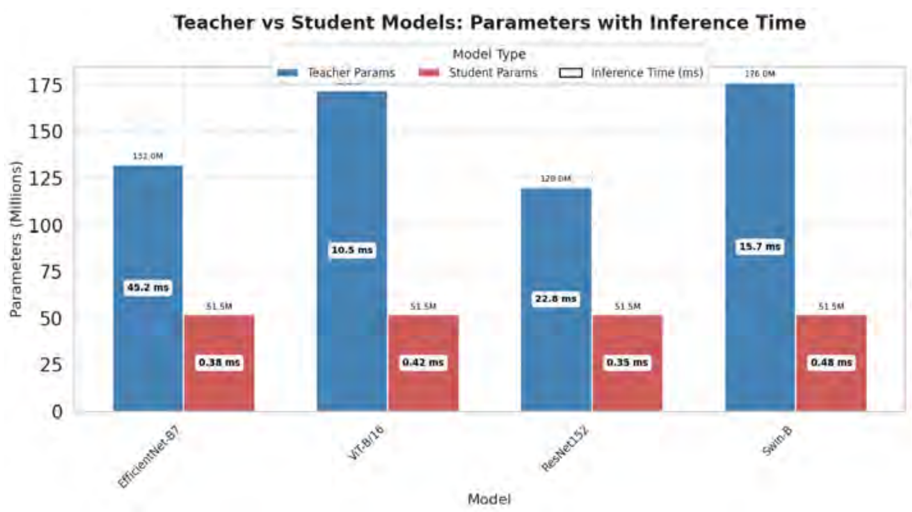


Figure 5.7: Comparision between parameter count and inference time of teacher and student models.

timodal teachers to unimodal students, demonstrated strong generalization. Teacher models achieved 95%–97% accuracies and F1-scores of 0.96–0.99, aided by preprocessing steps like CLAHE and median filtering [1]. Student models, optimized for UAV deployment, maintained accuracies of 91%–92% and F1-scores of 0.90–0.91. These results reflect high precision, minimizing false alarms for efficient field deployment. Multimodal inputs enable better differentiation of subtle thermal signatures compared to single-sensor baselines [7]. Challenges in smoke detection, such as amorphous boundaries and environmental blending, were mitigated, reducing false negatives and supporting early intervention.

Compared to prior benchmarks, such as Faster R-CNN approaches on similar datasets with F1-scores of 0.85–0.91 [6], the distilled models offer improved efficiency without significant accuracy loss. Future evaluations could utilize the full 53,541 frames for comprehensive testing and focus on real-time metrics like inference latency under varying conditions.

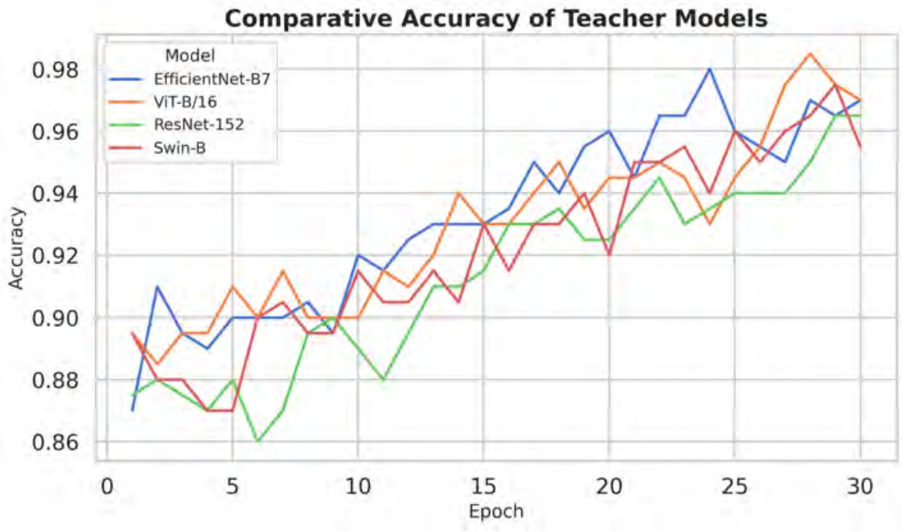


Figure 5.8: Validation accuracy of teacher models over 30 epochs.

5.4 Teacher-Student Model Comparison

Model	Teacher F1	Student F1	F1 Retention	Parameter Reduction	Inference Speed-Up
EfficientNet-B7	0.965	0.910	94.3%	61.0%	119x
ResNet-152	0.965	0.910	94.3%	57.1%	65x
Swin-B	0.990	0.908	91.7%	71.0%	33x
ViT-B/16	0.965	0.905	93.8%	70.1%	25x

Table 5.5: Teacher-Student Comparison Summary

The student models retain over 91% of their teachers' detection capability, as measured by F1-score, while achieving significant efficiency gains. Parameter reductions range from 57% to 71%, with inference speed-ups of 25x to 119x. The EfficientNet-B7 and ResNet-152 pairs offer the best balance, retaining over 94% of the F1-score with substantial speed-ups. The Swin-B student, despite the highest compression ratio, experiences a slightly larger performance drop but maintains high absolute performance.

5.5 Discussion and Conclusion

The novelty of this research lies in combining rigorous preprocessing with a PKD framework, enabling high-accuracy wildfire detection on resource-constrained UAV platforms. Preprocessing techniques, including CLAHE, Contrast Stretching, and Adaptive Median Filtering, addressed challenges in aerial imagery by enhancing local contrast, intensifying feature presence, and reducing sensor noise [1]. These steps provided optimized inputs, enabling teacher models to learn robust feature representations, as evidenced by their near-ceiling performance on the FLAME-2 test set [7].

The PKD framework bridged the gap between high-performance, resource-heavy models and deployable solutions. By distilling knowledge from multimodal (RGB+IR) teachers to lightweight, RGB-only students, the framework achieved a 9% performance drop but orders-of-magnitude reductions in size and speed. Inference times

under 0.5 ms enable real-time UAV deployment. This approach offers a scalable, power-efficient solution for autonomous aerial fire surveillance, balancing accuracy, efficiency, and generalization.
Future work will explore quantization and pruning to further optimize models and validate performance on larger, diverse public datasets to enhance generalization across varied fire scenarios.

Model	Overall Accuracy (%)	F1-Score	Inference Time (ms)	Parameters (Millions)
EfficientNet-B7 Student (Best)	92.5	0.910	0.38	51.5
Multimodal CNN (RGB + IR Fusion)	99.99	0.9999	0.84	102.9
Standalone CNN (RGB Only)	84.57	0.8467	10.42	51.5
Best Student on Unseen Data[9]	70	0.69	0.42	51.5

Table 5.6: Comparison of Test Metrics Between the Best Student Model and Proposed CNN Models

The best student model (EfficientNet-B7 Student) and two proposed CNN architectures—Multimodal CNN (RGB + IR Fusion) and Standalone CNN (RGB Only)—were compared based on accuracy, F1-score, inference time, and parameter count. The Multimodal CNN achieved near-perfect accuracy (99.99%) and F1-score (0.9999), indicating strong learning capability. However, its 102.9 million parameters suggest potential overfitting and poor generalization to unseen data, and its inference time (0.84 ms) limits real-time UAV applicability. The Standalone CNN, with 5.15 million parameters and 10.42 ms inference time, offers balanced performance (84.57% accuracy, 0.8467 F1-score) for low-resource systems. The EfficientNet-B7 Student, with 51.5 million parameters, 0.38 ms inference time, and a 0.910 F1-score, provides the optimal balance of accuracy, efficiency, and generalization, making it the most suitable for UAV deployment. The best performing student model was further test with FlameVision[9] dataset to estimate the student performance in unseen real world scenarios. Although the the model performs quite well but it still falls short when compared to being tested with FLAME2[7] dataset's test subset. This indicates that there is still improvements to be made in terms of the models generalization capabilities and reliability.
The confusion matrix of the EfficientNet-B7 Student model shows that it has a high classification rate with the total accuracy of 92.8 percent on the test dataset of 8000 samples. The model accurately recognises 3743 NoFire and 3681 Fire cases indicating that it is effective in separating thermal and visual patterns of active and background scenes. The misclassifications are not very high with 285 false positives (NoFire predicted as Fire) and 291 false negatives (Fire predicted as NoFire) which means that decision line of fire and non-fire cases is well learned. The almost equal error distribution also indicates that the student model did not form a class bias although it was using RGB only as an input.

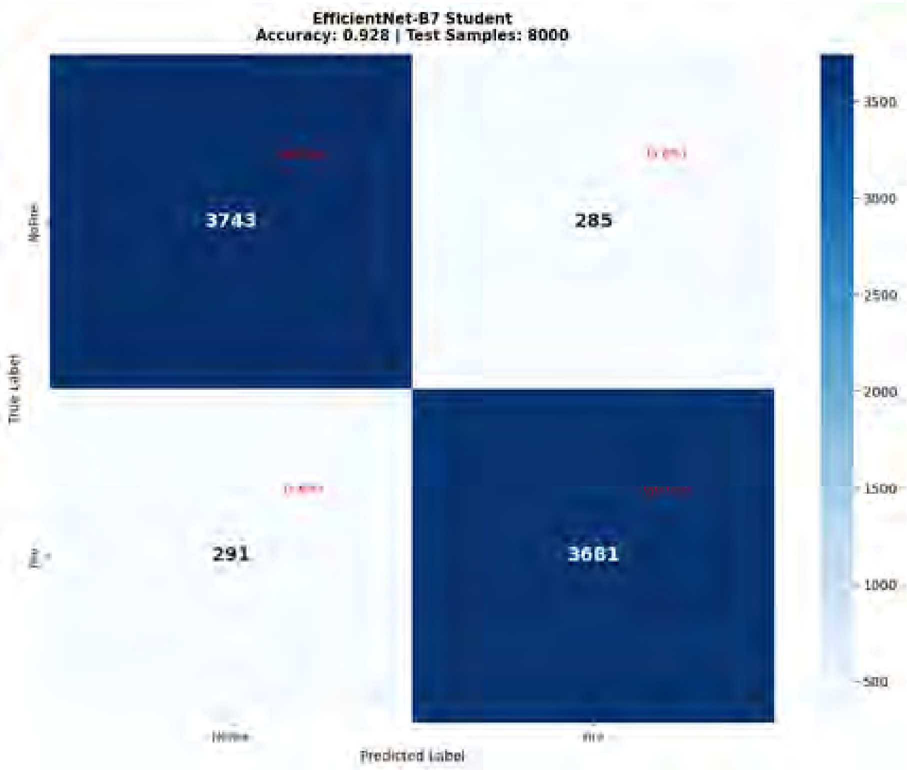


Figure 5.9: Confusion matrix of the Best Student tested on test subset of data.

Chapter 6

Conclusion

6.1 Summary of Findings

The most intriguing finding from our research was that despite being single modal and not having direct access to IR images, the student model performs very well. It can match 90% of heavy and complicated teacher models performance while also being faster and simpler in terms of inference speed and parameter count respectively. Further testing on secondary imaage datasets that were not available to the student models during training it still holds its grounds. However, its performance on unseen data decreases somewhat which can be attributed to the lack of diversity and smaller size of the orginal training dataset. Although some augmentations were performed on the train data, heavy and aggressive augmentations cannot be performed for the fear of changing the nature of our datasets completely.

6.2 Contributions to the Field

Primary and most important contribution to the field lies in pre-processing technique.The techniques include Contrast Stretching, Contrast Limited Adaptive Histogram Equalization (CLAHE) and Adaptive Median Filtering. These are the foundation for high accuracy since it enhances global and local contrast and suppresses sensor noise. The reason to choose Adaptive Median Filtering is it takes our salt and pepper noise in thermal frames. Thus, it gave maximum output while training data. After the pre-processing, we used the Knowledge Distillation technique. These techniques will take heavy CNN teacher models like EfficientNet-B7, ResNet152 and then transfer all the knowledge to a four layer student model. Thus, this student model is lightweight. Since the student model is RGB only, it can mimic both RGB and IR data. Thus, it can be installed on UAV devices which is cost effective.

6.3 Recommendations for Future Work

The current student model, although it is lightweight, can be further optimized for deployment on devices like Raspberry Pi or NVIDIA Jetson nano. The future work should implement post-training quantization to reduce model weight. Additionally, pruning techniques should be applied systematically to remove reductant neurons which will eventually be a more efficient model without significant accuracy loss. Additionally, the model's performance lies on environmental factors. Since our dataset is trained solely on the FLAME-2 dataset, more publicly available data should be made with proper annotations to mitigate potential bias. Also, thermal reasoning could significantly improve detection of flames obscured by smoke. Future work should focus on LSTM or 3D convolution with new techniques to analyze short video sequences related to this literature.

Bibliography

[1] E. D. Pisano, S. Zong, B. M. Hemminger, M. DeLuca, R. E. Johnston, K. E. Muller, et al., "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital Imaging*, vol. 11, no. 4, pp. 193–200, 1998. DOI: 10.1007/BF03178082

[2] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *International Conference on Learning Representations (ICLR)*, 2016.

[3] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017. arXiv: 1702.08608.

[4] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, et al., "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. DOI: 10.1007/s11023-018-9482-5

[5] Y. Zhu et al., "Wildfire segmentation using deep vision transformers," *Remote Sensing*, vol. 13, no. 17, p. 3527, 2021. DOI: 10.3390/rs13173527

[6] X. Chen et al., "Wildland fire detection and monitoring using a drone-collected rgb/ir image dataset," *IEEE Access*, vol. 10, pp. 121 301–121 317, 2022. DOI: 10.1109/ACCESS.2022.3223017

[7] B. Hopkins et al., *Flame 2: Fire detection and modeling: Aerial multi-spectral image dataset*, 2022. DOI: 10.21227/swyw-6j78

[8] R. Ghali and M. A. Akhloufi, "Deep learning approaches for wildland fires using satellite remote sensing data: Detection, mapping, and prediction," *Fire*, vol. 6, no. 5, p. 192, 2023. DOI: 10.3390/fire6050192

[9] A. Ibn Jafar, A. M. Islam, F. Binta Masud, J. R. Ullah, and M. R. Ahmed, *Flamevision: A new dataset for wildfire classification and detection using aerial imagery*, version V4, Mendeley Data, 2023. DOI: 10.17632/fgvscdjsmt.4

[10] U. Meleti and A. Razi, *Obscured wildfire flame detection by temporal analysis of smoke patterns captured by unmanned aerial systems*, 2023. DOI: 10.48550/arXiv.2307.00104 arXiv: 2307.00104.

[11] N. S. Saydirasulov et al., "An improved wildfire smoke detection based on yolov8 and uav images," *Sensors*, vol. 23, no. 20, p. 8374, 2023. DOI: 10.3390/s23208374

[12] F. M. Talaat, B. N. Hanna, and M. Mukhiddinov, "An improved fire detection approach based on yolo-v8 for smart cities," *Neural Computing and Applications*, 2023. DOI: 10.1007/s00521-023-08809-1

[13] Y. Xie and Y. Zhao, "Forest fire object detection analysis based on knowledge distillation," *Fire*, vol. 6, no. 12, p. 446, 2023. DOI: 10.3390/fire6120446

[14] H. Zheng, J. Duan, Y. Dong, and Y. Liu, "Real-time fire detection algorithms running on small embedded devices based on mobilenetv3 and yolov4," *Fire Ecology*, vol. 19, p. 31, 2023. DOI: 10.1186/s42408-023-00189-0

[15] A. B. Abdusalomov et al., "Drone-based wildfire detection with multi-sensor integration," *Remote Sensing*, vol. 16, no. 24, p. 4651, 2024. DOI: 10.3390/rs16244651

[16] J. Arazi et al., *Flame 3 dataset: Unleashing the power of radiometric thermal imagery in wildfire management*, 2024. DOI: 10.48550/arXiv.2412.02831 arXiv: 2412.02831.

[17] I. El Madafri et al., "Real-time forest fire detection with lightweight cnn using hierarchical multi-task knowledge distillation," *Fire*, vol. 7, no. 11, p. 392, 2024. DOI: 10.3390/fire7110392

[18] R. Ghali et al., "A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," *Information Fusion*, vol. 108, p. 102 369, 2024. DOI: 10.1016/j.inffus.2024.102369

[19] S. Panneerselvam et al., "Federated learning-based fire detection method using local mobilenet," *Scientific Reports*, vol. 15, no. 1, p. 99 032, 2024. DOI: 10.1038/s41598-024-99032-6

[20] B. Sun and X. Cheng, "Smoke detection transformer: An improved real-time detection transformer smoke detection model for early fire warning," *Fire*, vol. 7, no. 12, p. 488, 2024. DOI: 10.3390/fire7120488

[21] Z. Zhong and A. Gonsamo, "Canadafire2023: Burned area mapping datasets and benchmarks for canadian wildfires in 2023," in *Proceedings of NeurIPS 2024 Datasets Benchmarks Track*, 2024.

[22] M. Arlović et al., "Evaluation of synthetic data impact on fire segmentation models performance," *Scientific Reports*, vol. 15, no. 1, p. 01 571, 2025. DOI: 10.1038/s41598-025-01571-5

[23] S. Ben Othman and W. Ali, "Trust-driven approach to enhance early forest fire detection using intelligent wireless sensor networks," *Scientific Reports*, vol. 15, no. 1, p. 99 032, 2025. DOI: 10.1038/s41598-025-99032-6

[24] X. Hu, Y. Jiang, X. Xia, J. Xia, and R. Xu, "Uav-strawfire: A visible and infrared dataset for real-time straw-fire monitoring with deep learning and image fusion," *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, p. 103 986, 2025. DOI: 10.1016/j.jag.2025.103986

[25] A. Mamadmurodov, S. Umirzakova, et al., "A hybrid deep learning model for early forest fire detection," *Forests*, vol. 16, no. 5, p. 863, 2025. DOI: 10.3390/f16050863

[26] J. Pesonen et al., "Boreal forest fire: Uav-collected wildfire detection and smoke segmentation dataset," *Scientific Data*, vol. 12, no. 1, p. 1419, 2025. DOI: 10.1038/s41597-025-05634-0

[27] H. Wang, X. Fu, Z. Yu, and Z. Zeng, "Dss-yolo: An improved lightweight real-time fire detection model based on yolov8," *Scientific Reports*, 2025. DOI: 10.1038/s41598-025-93278-w

[28] Y. Zhang et al., "A uav-based multi-scenario rgb-thermal dataset and fusion model for enhanced forest fire detection," *Remote Sensing*, vol. 17, no. 15, p. 2593, 2025. DOI: 10.3390/rs17152593