Journal of Computing and Smart Ecosystem
Universitas Muhammadiyah Semarang
Vol. 2 No.1, 2026: 67-83

J-CaSE
e-ISSN 3110-5777

Comparative Evaluation Of Efficientnet-B0 And Mobilenetv3 For Multi-Class Fire Image Classification With Grad-Cam Interpretation

Muhammad Fabian Rizky Fatah[1], Ravicenna Mahardhika[1], Muhammad Falah Altairgunna[1], Ricardus Anggi Pramunendar[1]
[1] Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia

***Corresponding author**: *111202315301@mhs.dinus.ac.id*

Article Info:
Received:
Accepted:
Available Online:

Keywords:
EfficientNet-B0; MobileNetV3; Transfer Learning; Fire Classification; Grad-CAM; Image Classification
.

Abstract: Fire detection using image-based computer vision is a promising alternative to conventional sensor systems, which are often limited by detection range and false alarms. This study presents a comparative evaluation of two lightweight deep learning architectures, EfficientNet-B0 and MobileNetV3 Small, for multi-class fire image classification across four categories: Urban Fire, Wild Fire, Non-Damage Building, and Non-Damage Wildlife. Both models were trained using transfer learning with ImageNet pre-trained weights on a curated dataset of 1,933 images with selective undersampling to address class imbalance. Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to interpret model decisions and validate the learned visual features. EfficientNet-B0 achieved the highest overall accuracy of 93.8% with a macro F1-score of 0.925, while MobileNetV3 Small reached 91.0% validation accuracy with a macro F1-score of 0.900. Grad-CAM visualizations confirmed that EfficientNet-B0 develops more localized feature activations focused on flame-colored regions, while MobileNetV3 exhibited broader contextual activations. The results demonstrate that lightweight transfer learning models combined with Grad-CAM provide both high classification performance and meaningful interpretability for fire detection applications.

1. INTRODUCTION

Fire is one of the most frequently occurring disasters and causes substantial losses both in material damage and human casualties. In Indonesia, fires in densely populated residential areas, multi-story buildings, as well as forest and peatland areas have become a recurring threat every year [1]. Delayed fire detection, even at a minor scale, can trigger uncontrolled escalation and worsen the overall impact of the disaster.

Conventional sensor-based fire detection approaches, such as smoke detectors and heat sensors, have significant limitations, including dependence on the distance from the fire source, relatively slow response times, and the potential for false detections due to environmental conditions such as dust or vapor [2]. With the advancement of artificial intelligence technology, particularly deep learning-based computer vision, fire detection and classification through image analysis has become a promising alternative, as it is capable of processing visual information automatically and adapting more flexibly to various environmental conditions [3],

[DOI: 10.14710/JCASE. vol2.iss1.1129]
OPEN ACCESS
CC BY SA

[4]. Foundational architectures such as ResNet [5], Inception [6], InceptionV3 [7] and DenseNet [8] have laid the groundwork for modern CNN development, enabling subsequent lightweight variants to achieve competitive performance with significantly reduced computational overhead.

Several studies have demonstrated the effectiveness of deep learning approaches for fire detection. Faris et al.raha [2] implemented MobileNetV2 for a real-time image-based house fire detection system. At an optimal distance of 3 meters for paper burning, the system achieved an accuracy of 99.75% with an average response time ranging from 74.5 ms to 117.1 ms depending on internet connection quality. Nguyen et al. [3] proposed a lightweight CNN architecture combined with an attention mechanism and inception module, achieving accuracy above 96% across six different datasets and deployable on resource-constrained devices such as the Jetson Nano at 14.27 FPS. Sozol et al. [1] developed a YOLOv5-based indoor fire detection model optimized using a genetic algorithm, achieving a mAP@0.5 of 92.1% on a dataset of 5,000 indoor fire images.

Class imbalance is a common challenge in fire datasets. Bahtiar et al. [9] addressed this issue using a class weighting approach with EfficientNet-B1 for multi-class forest fire classification, achieving 90% accuracy on 7,331 images distributed across four classes including a smoke class, with a training time of only 14 minutes and 45 seconds. In the medical imaging domain, Bintang & Imaduddin [10] compared four CNN architectures using transfer learning for diabetic retinopathy detection, where Xception and DenseNet201 achieved accuracy and F1-scores up to 96%, demonstrating the effectiveness of transfer learning on limited domain-specific datasets. In the environmental imaging domain, Muhammad et al. [11] developed a coral reef classification model using a dataset of 923 images from Flickr. Their results showed that a ResNet model trained from scratch outperformed pre-trained models, indicating that an appropriate training strategy plays a role no less important than the size of the dataset.

Although these studies demonstrate high performance, deep learning models are generally black-box in nature, making their decision-making processes difficult to understand [4], [12]. In critical systems such as fire detection, the absence of transparency can hinder user trust in the system. Grad-CAM (Gradient-weighted Class Activation Mapping) addresses this by generating visual heatmaps that highlight which areas of an image most influence the model's decision [13].

Apostolopoulos et al. [14] evaluated 17 different CNN architectures for smoke and fire detection using 53,585 images, finding that Xception [15] produced the highest accuracy, and that Grad-CAM++ and LIME successfully verified that the model focused on critical image locations containing smoke or fire. Alam et al. [16] developed FireNet-CNN, a real-time forest fire detection model integrating Grad-CAM and saliency maps. With 2.75 million parameters and a size of 10.58 MB, this model achieved an accuracy of 99.05%, precision of 99.41%, and recall of 98.28%, with an inference speed of 0.95 seconds per image, making it suitable for deployment on drones and embedded devices. Yar et al. [17] integrated modified channel and spatial attention mechanisms into EfficientNetV2B0 and employed Grad-CAM, guided backpropagation, and guided Grad-CAM to visualize salient areas identified by the model across various benchmark datasets.

CNN-based smoke detection research by Khan et al. [18] using the Smoke Detection Dataset also showed that despite an overall accuracy of 0.994, precision of 0.69 and recall of

[DOI: 10.14710/JCASE. vol2.iss1.1129]

0.78 on the smoke class remained lower, indicating that classifying visually hazardous conditions is inherently more challenging than normal conditions. This finding is relevant to the broader context of fire image classification research.

Based on the literature review, a research gap exists: few studies have simultaneously conducted a comparative evaluation of lightweight deep learning models for multi-class fire classification alongside model interpretation using Grad-CAM within a single pipeline directed at understanding model behavior. Therefore, this study aims to conduct a comparative evaluation of EfficientNet-B0 and MobileNetV3 Small using a transfer learning approach on a four-class fire dataset with an intelligent data curation strategy, and to apply Grad-CAM to interpret and compare the feature representation behavior of both models.

The main contributions of this study include: (1) a comparative evaluation of EfficientNet-B0 and MobileNetV3 Small on a four-class fire dataset with a selective undersampling strategy for handling class imbalance; and (2) model interpretability analysis using Grad-CAM on correctly predicted cases (True Positive) and prediction failures (False Negative) to understand the feature representation behavior of each model.

2. METHODOLOGY

This study was conducted through several stages. First, the Disaster Images Dataset was collected from Kaggle and data curation was performed using a selective undersampling strategy to balance the distribution of four classes. Subsequently, data preprocessing and augmentation were carried out, the dataset was split at an 80:20 ratio, two models were trained comparatively using a transfer learning approach, model performance was evaluated using various metrics, and Grad-CAM was applied for model interpretation. The final stage involves a comparative analysis of both models based on quantitative metrics and Grad-CAM interpretability results.

[DOI: 10.14710/JCASE. vol2.iss1.1129]

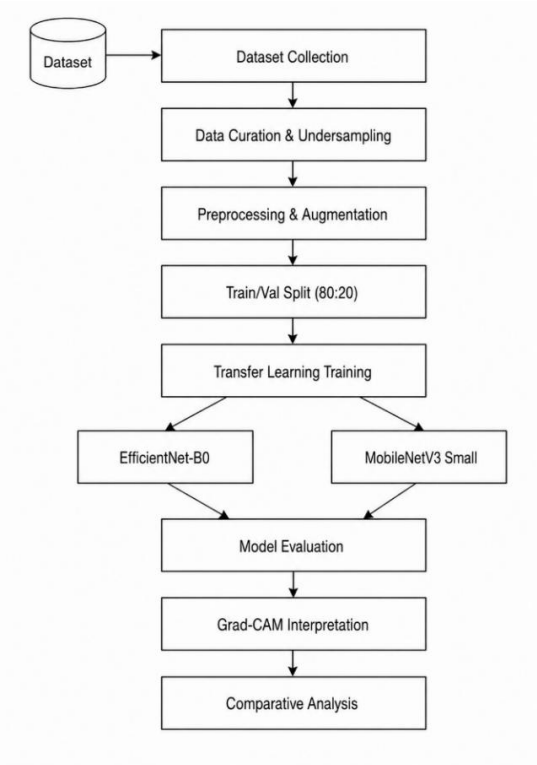


Fig 1. Research Methodology Flowchart

2.1 Dataset

The dataset used in this study was sourced from the Kaggle platform under the name Disaster Images Dataset, curated by user varpit94. The original dataset consists of various natural disaster classes; however, only four classes were selected in accordance with the research focus: Urban_Fire, Wild_Fire, Non_Damage_Building, and Non_Damage_Wildlife. The inclusion of Non_Damage classes was intended to train the model to recognize normal conditions, enabling it to more accurately distinguish hazardous conditions from safe ones.

Most images in this dataset are in JPEG format with varying resolutions, sourced from surveillance footage, field photographs, and aerial imagery captured under diverse lighting conditions and angles. This diversity of sources makes the dataset sufficiently representative for real-world fire detection scenarios, despite the limited total volume of data. Figure 2 illustrates representative sample images from each of the four classes used in this study.

[DOI: 10.14710/JCASE. vol2.iss1.1129]

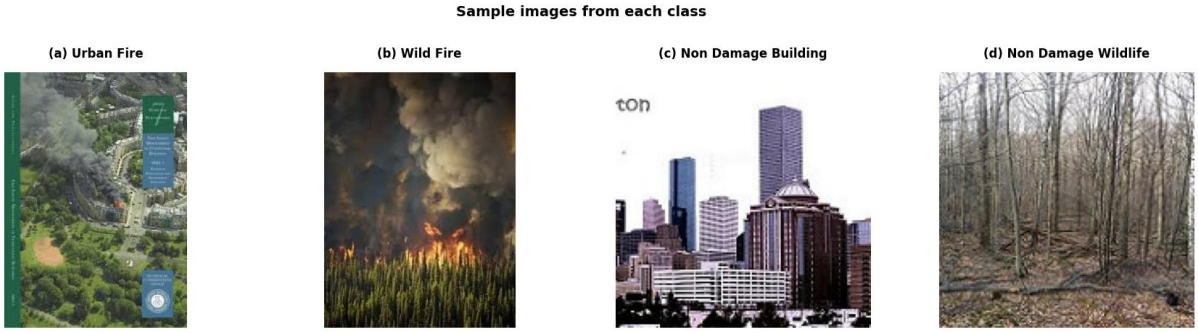


Fig 2. Sample images from each class: (a) Urban_Fire, (b) Wild_Fire, (c) Non_Damage_Building, (d) Non_Damage_Wildlife

An intelligent data curation strategy (smart data restructuring) was applied to address class imbalance. All images in both fire classes were retained — specifically 419 Urban_Fire and 514 Wild_Fire images — while the Non_Damage classes were capped at a maximum of 500 images per class through random undersampling. The final dataset totals 1,933 images. Table 1 presents the dataset distribution along with the train-validation split.

Table 1. Dataset Distribution and Split

Class	Total Images	Train Split	Val Split
Urban_Fire	419	335	84
Wild_Fire	514	411	103
Non_Damage_Building	500	400	100
Non_Damage_Wildlife	500	400	100
Total	**1,933**	**1,546**	**387**

Normalization was performed using a mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225] in accordance with ImageNet dataset statistics [19]. Data augmentation was applied to the training set, including random horizontal flipping, random rotation of ±15°, and color jitter on brightness, contrast, and saturation at a factor of 0.2. Augmentation was intended to improve model robustness against lighting variation commonly encountered in field fire imagery [20]. All images were resized to 224×224 pixels to meet the standard input requirements of both model architectures.

2.2 Model Architecture

Two lightweight deep learning architectures were comparatively evaluated in this study. Both were selected because they are designed to balance performance and computational efficiency, which is relevant for deployment scenarios on resource-constrained devices [21].

EfficientNet-B0 is the smallest variant of the EfficientNet family introduced by Tan and Le [21]. The primary advantage of EfficientNet lies in its compound scaling strategy, which scales network dimensions — depth, width, and resolution — proportionally and

[DOI: 10.14710/JCASE. vol2.iss1.1129]

simultaneously using coefficients determined through neural architecture search. This approach yields a model that is empirically more efficient than the single-dimension scaling strategies commonly used previously. EfficientNet-B0 employs MBConv (Mobile Inverted Bottleneck Convolution) blocks that combine depthwise separable convolutions with a squeeze-and-excitation mechanism for adaptive channel feature weighting. With approximately 5.3 million parameters, EfficientNet-B0 is classified as a lightweight model capable of achieving competitive performance on the ImageNet benchmark.

MobileNetV3 Small is the result of further development in the MobileNet series, introduced by Howard et al. [22]. This architecture is specifically designed for deployment on mobile and edge devices by optimizing performance per computational operation. MobileNetV3 integrates three key innovations: (1) the use of depthwise separable convolutions to drastically reduce the number of parameters and computations compared to standard convolutions; (2) the integration of a squeeze-and-excitation mechanism for adaptive channel feature recalibration; and (3) the use of the hard-swish activation function, which is computationally more efficient than standard swish. The Small variant contains approximately 2.5 million parameters, making it one of the lightest CNN architectures available for image classification.

Transfer learning from ImageNet pre-trained weights [19] was used as the starting point for training both models, following the seminal demonstration by Krizhevsky et al. [23] that deep CNNs trained on large-scale image datasets can learn highly transferable visual representations. This approach has been proven effective in improving model performance on limited domain-specific datasets, as pre-trained weights contain general visual feature representations that can be transferred to new classification tasks [24]. Both models were configured as end-to-end trainable with an output class count of 4 corresponding to the dataset classes. Implementation used the timm (PyTorch Image Models) library, which provides standard implementations of both architectures along with ImageNet pre-trained weights.

2.3 Training Configuration

The Adam optimizer [25] was used with an initial learning rate of 0.001 and a weight decay of 1×10^{-4} for regularization to prevent overfitting. The loss function used was Cross-Entropy Loss, which is appropriate for multi-class classification tasks. The learning rate schedule used a Cosine Annealing LR Scheduler [26] with T_max = 15 to smoothly decrease the learning rate following a cosine curve throughout the training process, helping the model achieve more stable convergence. Both models were trained for 15 epochs with a checkpointing mechanism that saves the best weights based on the highest validation accuracy. Training was conducted using an NVIDIA GPU on Google Colab. All training configurations are summarized in Table 2.

[DOI: 10.14710/JCASE. vol2.iss1.1129]

Table 2. Model Training Configuration

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Weight Decay	1e-4
Scheduler	CosineAnnealingLR
T_max	15
Epochs	15
Batch Size	32
Input Size	224 × 224 pixels
Loss Function	CrossEntropyLoss
Framework	PyTorch + timm

The hyperparameter values used in this study were determined through manual tuning; future work may benefit from automated hyperparameter optimization frameworks such as Optuna [27] to systematically explore a broader search space and potentially improve model performance.

2.4 Model Evaluation

Model performance was evaluated on the validation set using precision, recall, F1-score, and overall accuracy computed both per class and as macro average and weighted average. A confusion matrix was used to visualize classification error patterns in detail for each class pair. Training loss and validation loss curves per epoch were also analyzed to understand training dynamics and detect potential overfitting or model instability.

2.5 Grad-CAM Application

Grad-CAM (Gradient-weighted Class Activation Mapping) [13] was applied using the pytorch-grad-cam library to generate interpretability visualizations on both models. Grad-CAM works by computing the gradient of the target class score with respect to the activation maps of a designated convolutional layer. These gradients are spatially aggregated to produce importance weights for each feature map, which are then combined and processed through a ReLU function to generate a non-negative heatmap. The heatmap is overlaid on the original image using a jet colormap, where red areas indicate the regions most influential to the model's prediction.

For EfficientNet-B0, the target layer used was conv_head, the last convolutional layer before global average pooling. For MobileNetV3, the target layer used was the last convolutional block (blocks[-1]). Grad-CAM analysis was performed on True Positive image samples (correct predictions on fire classes) and False Negative samples (detection failures) to

[DOI: 10.14710/JCASE. vol2.iss1.1129]

provide a comprehensive understanding of each model's feature representation behavior under different prediction conditions.

While this study employs the standard Grad-CAM method, its improved variant Grad-CAM++ [28] offers more precise localization particularly for images containing multiple object instances, and may be considered in future work for more granular activation analysis.

3. RESULTS AND DISCUSSION

3.1 Dataset Curation and Distribution

The dataset curation process yielded a total of 1,933 images distributed across four classes. The Wild_Fire class has the largest count with 514 images, followed by Non_Damage_Building and Non_Damage_Wildlife with 500 images each, and Urban_Fire with 419 images. The strategy of capping the Non_Damage classes at 500 images successfully maintained the ratio between fire and non-fire classes within a relatively balanced range, with approximately 48% fire images and 52% non-fire images of the total dataset. Although the total dataset is relatively limited, the selective undersampling strategy successfully maintained class balance, and the use of ImageNet pre-trained weights through transfer learning mitigates the risk of overfitting on small datasets, as demonstrated by the stable validation accuracy achieved by both models.

3.2 Training Performance

Both models exhibited different training dynamics. EfficientNet-B0 demonstrated more stable convergence, with its best validation accuracy of 93.8% achieved at epoch 3 of the second training cycle. In contrast, MobileNetV3 showed a more fluctuating training pattern, particularly in the middle and later epochs where the validation loss experienced significant spikes, indicating a higher sensitivity to learning rate changes during the final phase of CosineAnnealingLR.

The model checkpointing mechanism proved crucial in this study. For MobileNetV3, the best weights were saved at epoch 4 with a validation accuracy of 91.5%, well before the model experienced instability in subsequent epochs. This underscores the importance of saving the best weights during training rather than relying solely on the weights at the final epoch — a finding consistent with best practices in deep learning model training [4].

[DOI: 10.14710/JCASE. vol2.iss1.1129]

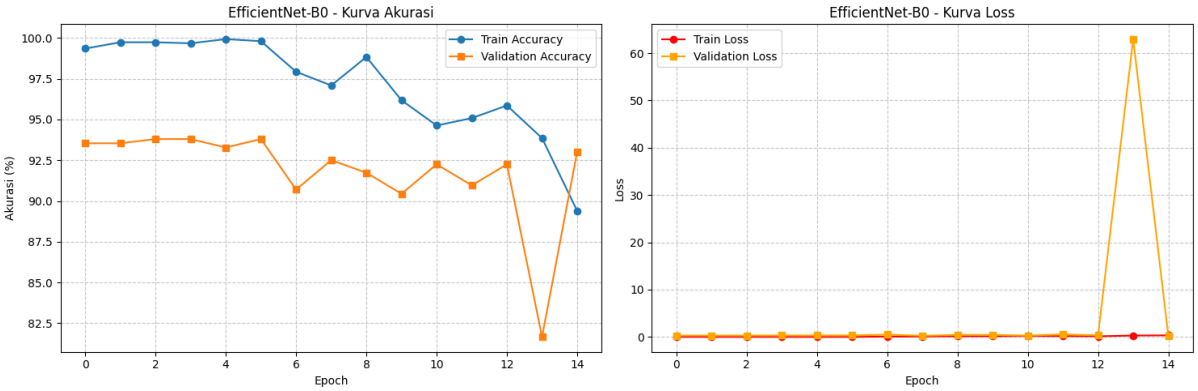


Fig 3. Accuracy and Loss Curve of EfficientNet-B0

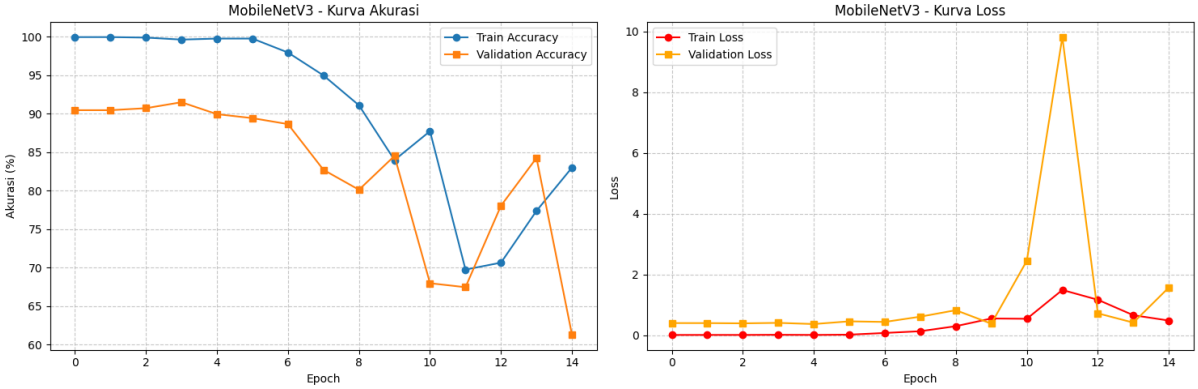


Fig 4. Accuracy and Loss Curve of MobileNetV3 Small

3.3 Classification Performance Evaluation

Table 3 presents the comparative classification report of both models on the validation set, including precision, recall, and F1-score per class along with overall averages.

Table 3. Comparative Classification Report

Kelas	**EfficientNet-B0**			**MobileNetV3 Small**		
	Presisi	Recall	F1	Presisi	Recall	F1
Non_Damage_Building	0,97	1,00	0,99	0,95	0,99	0,97
Non_Damage_Wildlife	1,00	0,98	0,99	0,98	1,00	0,99
Urban_Fire	0,85	0,82	0,84	0,83	0,71	0,77

[DOI: 10.14710/JCASE. vol2.iss1.1129]

Kelas	EfficientNet-B0			MobileNetV3 Small		
Wild_Fire	0,86	0,88	0,87	0,83	0,88	0,85
Akurasi	**0,938**			**0,910**		
Macro Avg	0,92	0,92	0,92	0,90	0,90	0,90
Weighted Avg	0,92	0,93	0,92	0,90	0,90	0,90

EfficientNet-B0 achieved an overall accuracy of 93.8% with a macro F1-score of 0.925. The highest performance was recorded on the Non_Damage classes with an F1-score of 0.99 for both subclasses, indicating that the model is highly effective at identifying normal conditions. For the fire classes, Urban_Fire obtained an F1-score of 0.84 with precision of 0.85 and recall of 0.82, while Wild_Fire achieved an F1-score of 0.87 with precision of 0.86 and recall of 0.88.

MobileNetV3 Small achieved its best validation accuracy of 91.0% with a macro F1-score of 0.900. Performance on the Non_Damage classes was also high, yet lower than EfficientNet-B0 on the fire classes, particularly Urban_Fire, which only achieved an F1-score of 0.77 with a recall of 0.71. The largest performance gap between the two models occurred in the Urban_Fire class, with a recall difference of 0.82 vs. 0.71. This is likely attributable to the visually diverse characteristics of Urban_Fire images — encompassing fires in various urban settings with complex artificial lighting — which require greater feature representation capacity, as provided by EfficientNet-B0 through its compound scaling strategy [21].

Lower performance on fire classes compared to Non_Damage classes is consistent with findings from similar studies. Khan et al. [18] also reported that despite a very high overall accuracy of 0.994, precision (0.69) and recall (0.78) on the smoke class remained lower than on the non-smoke class, indicating that classifying visually hazardous conditions is inherently more challenging due to greater appearance variability.

[DOI: 10.14710/JCASE. vol2.iss1.1129]

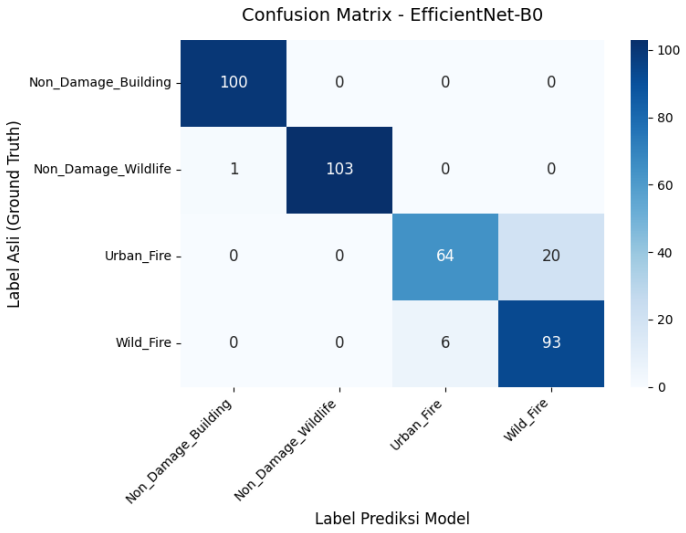


Fig 5. Confusion Matrix of EfficientNet-B0

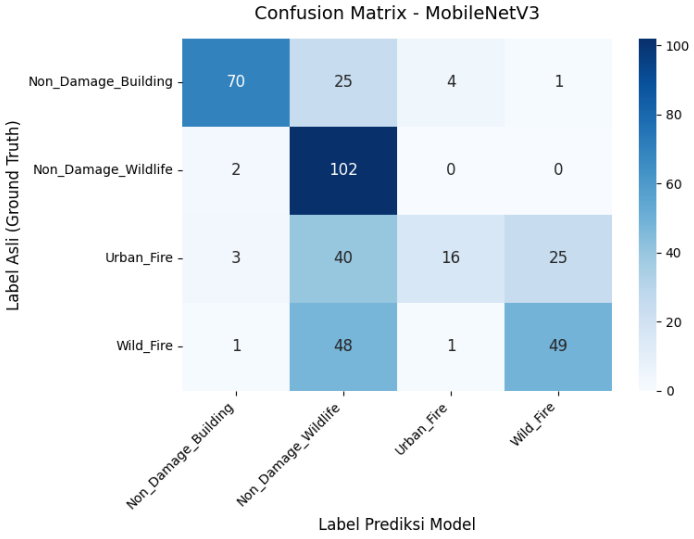


Fig 6. Confusion Matrix of MobileNetV3 Small

3.4 Grad-CAM Analysis

Grad-CAM analysis was focused on the Urban_Fire class as the class with the highest error rate across both models. In True Positive cases, EfficientNet-B0's Grad-CAM showed activations concentrated on flame regions and smoke sources, producing a clear and localized red-orange heatmap. This pattern confirms that the model has successfully learned semantically meaningful feature representations — specifically, the visual characteristics of fire in the form of orange-red coloring with high luminance and fluctuating contours. This finding aligns with

[DOI: 10.14710/JCASE. vol2.iss1.1129]

the results of Apostolopoulos et al. [14], who used Grad-CAM++ and LIME to verify feature localization on fire and smoke regions in fire imagery.

MobileNetV3's Grad-CAM on the same True Positive cases exhibited a more diffuse activation pattern compared to EfficientNet-B0. Although the fire region still received high activation, the broader heatmap distribution indicates that MobileNetV3 also considers contextual features surrounding the fire, such as sky color and nearby building structures. This more diffuse activation pattern directly correlates with the lower recall on the Urban_Fire class — the model tends to be uncertain when contextual features are inconsistent with the patterns learned during training.

In False Negative cases, Grad-CAM reveals that both models struggle when the fire source is small within the frame or when the fire is obscured by thick smoke that dominates the image. Under these conditions, activation is spread across the smoke region and background without a clear focus on the actual fire area. This finding is consistent with the results of Alam et al. [16], who identified dominant smoke presence as one of the primary factors contributing to model failure in fire detection. These findings provide a clear direction for dataset improvement, namely by adding more samples of fires with small flames and smoke-dominated fires.

Overall, Grad-CAM analysis successfully validated that EfficientNet-B0 develops more localized and semantically meaningful feature representations compared to MobileNetV3, which directly correlates with its higher classification performance. These results support the findings of Yar et al. [17] that the quality of feature localization indicated by Grad-CAM is closely related to the model's ability to understand relevant visual characteristics in fire imagery.

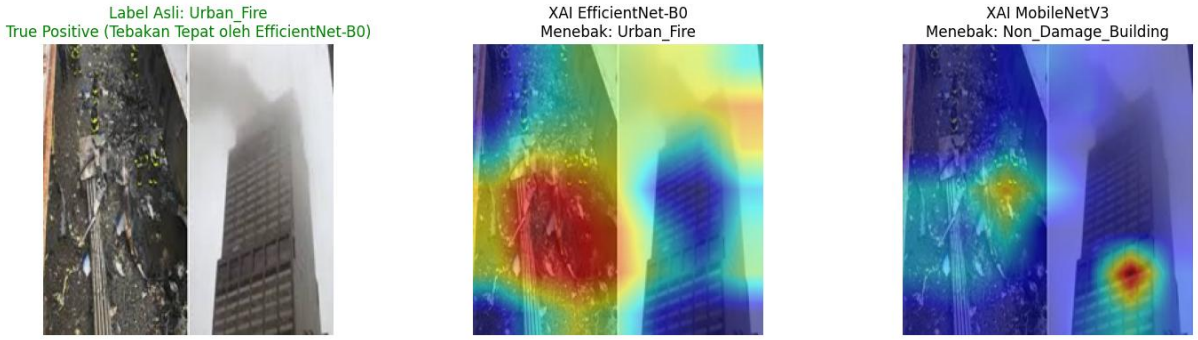


Fig 7. Grad-CAM of EfficientNet-B0 True Positive

[DOI: 10.14710/JCASE. vol2.iss1.1129]

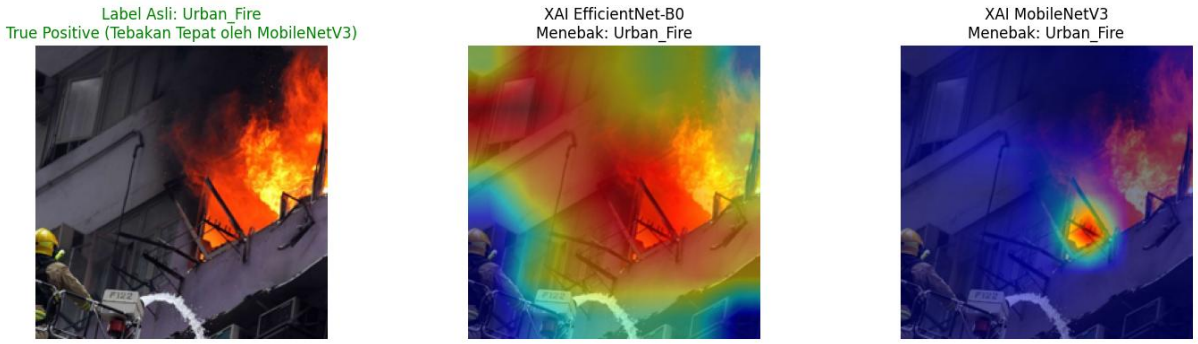


Fig 8. Grad-CAM MobileNetV3 True Positive

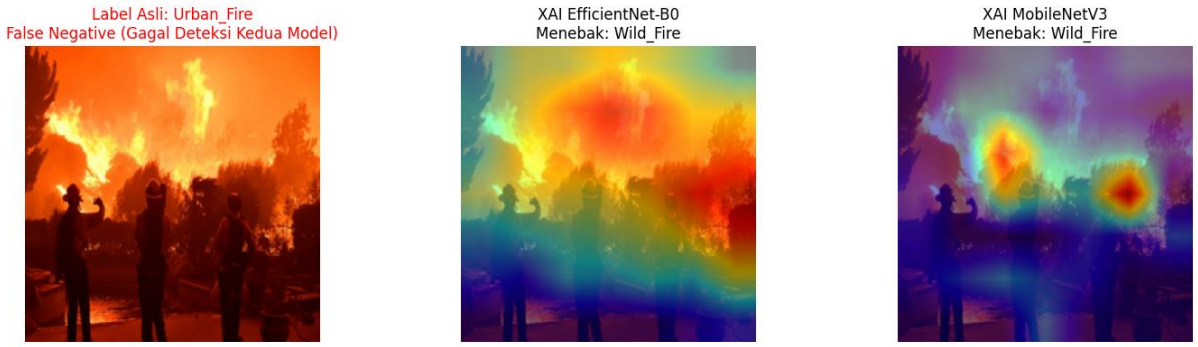


Fig 9. Grad-CAM on False Negative Cases

3.5 Comparison with Related Studies

Direct accuracy comparisons with previous studies cannot be made due to differences in the datasets used across studies. Table 4 presents a summary of related studies as context for positioning this research within the broader landscape of deep learning-based fire classification research.

Table 4. Review of Related Studies

Study	Model	Classes	Accuracy	XAI
Faris et al. [2]	MobileNetV2	2	99.75%	—
Nguyen et al. [3]	Lightweight CNN	2	>96%	—
Sozol et al. [1]	YOLOv5 (HPO)	2	mAP 92.1%	✓
Bahtiar et al. [9]	EfficientNet-B1	4	90%	—
Apostolopoulos et al. [14]	Xception + XAI	2	High	✓

[DOI: 10.14710/JCASE. vol2.iss1.1129]

Alam et al. [16]	FireNet-CNN	2	99.05%	✓
Yar et al. [17]	EFNet-CSM	2	—	✓
Proposed	EfficientNet-B0	4	93.8%	✓
Proposed	MobileNetV3 Small	4	91.0%	✓

Several distinguishing points of this study can be identified from the table. First, the majority of prior studies used binary classification (two classes), while this study employs four classes encompassing two types of fire and two types of normal conditions. Multi-class classification is inherently more challenging because the model must distinguish more diverse visual patterns, making the 93.8% accuracy achieved by EfficientNet-B0 a competitive result in this context [9].

Second, in terms of dataset size, this study uses 1,933 images, which is relatively smaller than Bahtiar et al. [9] who used 7,331 images in a similar four-class configuration. Nevertheless, the intelligent data curation strategy through selective undersampling applied in this study produced performance that surpassed that study (93.8% vs. 90%), indicating that the quality and balance of data distribution can compensate for quantitative limitations [20]. This also aligns with the findings of Muhammad et al. [11], who demonstrated that an appropriate training strategy plays a role no less important than dataset size alone.

Third, from an interpretability perspective, this study — together with Apostolopoulos et al. [14], Alam et al. [16], and Yar et al. [17] — belongs to the group of studies that integrate model interpretability techniques as part of the evaluation pipeline, rather than merely reporting performance metrics. This approach is increasingly relevant given the growing demands for transparency and accountability in AI systems [29], particularly in the domain of critical applications such as disaster detection.

4. CONCLUSION

This study conducted a comparative evaluation of EfficientNet-B0 and MobileNetV3 Small for multi-class fire image classification using a transfer learning approach, complemented by model interpretation using Grad-CAM. Several key conclusions can be drawn from this research.

EfficientNet-B0 consistently demonstrated superior performance compared to MobileNetV3 Small across all evaluation metrics. EfficientNet-B0 achieved an accuracy of 93.8% with a macro F1-score of 0.925, while MobileNetV3 reached its best accuracy of 91.0% with a macro F1-score of 0.900. The most significant performance gap occurred in the Urban_Fire class, which exhibits the highest visual variability in the dataset. This aligns with the compound scaling characteristics of EfficientNet-B0, which provides greater feature representation capacity compared to MobileNetV3's depthwise separable architecture [21].

[DOI: 10.14710/JCASE. vol2.iss1.1129]

The application of Grad-CAM successfully validated that EfficientNet-B0 develops more localized and semantically meaningful feature representations, with heatmaps that consistently focus on flame regions and fire-related visual characteristics. MobileNetV3 Small exhibited more diffuse activation patterns that relied on contextual environmental features, which correlated with lower performance, particularly in Urban_Fire recall. Analysis of False Negative cases revealed that both models struggle with fire images containing small flames or those dominated by smoke, providing a clear direction for dataset improvement in future research.

The intelligent data curation strategy through selective undersampling of majority classes proved effective in maintaining class balance and producing competitive performance despite a limited total dataset of 1,933 images. The model checkpointing mechanism also proved crucial, particularly for MobileNetV3, which exhibited instability in the final training phase due to the dynamics of CosineAnnealingLR [26].

For future research, several recommendations can be provided: (1) dataset expansion with incremental fire samples and smoke-dominated fire images to improve recall on fire classes; (2) exploration of newer model architectures such as EfficientNetV2 or Vision Transformer-based models [30], [31] for multi-class fire image classification; and (3) evaluation of the models under real-world testing conditions including more extreme weather and lighting variation.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Informatics Engineering Study Program, Universitas Dian Nuswantoro, for the facility support and guidance provided throughout the course of this research. The authors also extend their appreciation to the provider of the Disaster Images Dataset on the Kaggle platform, which made this research possible.

REFERENCES

[1] M. S. S. Sozol, M. R. H. Mondal, and A. H. Thamrin, “Indoor fire and smoke detection based on optimized YOLOv5,” *PLoS One*, vol. 20, Apr. 2025, doi: 10.1371/journal.pone.0322052.

[2] M. Faris, E. Ariyanto, and Y. A. S. Yudo, “IMPROVED REAL-TIME HOUSE FIRE DETECTION SYSTEM PERFORMANCE WITH IMAGE CLASSIFICATION USING MOBILENETV2 MODEL,” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, pp. 656–663, May 2023, doi: 10.29100/jipi.v8i2.3803.

[3] D. L. Nguyen, M. D. Putro, and K. H. Jo, “Lightweight Convolutional Neural Network for Fire Classification in Surveillance System,” *IEEE Access*, vol. 11, pp. 101604–101615, 2023, doi: 10.1109/ACCESS.2023.3305455.

[4] D. Bhatt *et al.*, “CNN variants for computer vision: History, architecture, application, challenges and future scope,” Oct. 2021, *MDPI*. doi: 10.3390/electronics10202470.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[6] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Oct. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[9] A. Bahtiar, M. I. P. Hutomo, A. Widiyanto, and S. Khomsah, "Class Weighting Approach for Handling Imbalanced Data on Forest Fire Classification Using EfficientNet-B1," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, pp. 63–73, Jan. 2025, doi: 10.14421/jiska.2025.10.1.63-73.

[10] Y. K. Bintang and H. Imaduddin, "PENGEMBANGAN MODEL DEEP LEARNING UNTUK DETEKSI RETINOPATI DIABETIK MENGGUNAKAN METODE TRANSFER LEARNING," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, pp. 1442–1455, Aug. 2024, doi: 10.29100/jipi.v9i3.5588.

[11] F. Muhammad, A. B. Elfandra, I. P. Amin, and A. F. Wicaksono, "Pengembangan Model untuk Mendeteksi Kerusakan pada Terumbu Karang dengan Klasifikasi Citra," Jan. 2026, [Online]. Available: https://arxiv.org/abs/2308.04337

[12] M. Mehmood, A. Shahzad, B. Zafar, A. Shabbir, and N. Ali, "Remote Sensing Image Classification: A Comprehensive Review and Applications," 2022, *Hindawi Limited*. doi: 10.1155/2022/5880959.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[14] I. D. Apostolopoulos, I. Athanasoula, M. Tzani, and P. P. Groumpos, "An Explainable Deep Learning Framework for Detecting and Localising Smoke and Fire Incidents: Evaluation of Grad-CAM++ and LIME," *Mach. Learn. Knowl. Extr.*, vol. 4, pp. 1124–1135, Dec. 2022, doi: 10.3390/make4040057.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[16] G. M. I. Alam, N. Tasnia, T. Biswas, M. J. Hossen, S. A. Tanim, and M. S. U. Miah, "Real-Time Detection of Forest Fires Using FireNet-CNN and Explainable AI Techniques," *IEEE Access*, vol. 13, pp. 51150–51181, 2025, doi: 10.1109/ACCESS.2025.3552352.

[17] H. Yar, F. U. M. Ullah, Z. A. Khan, M. J. Kim, and S. W. Baik, "EFNet-CSM: EfficientNet with a modified attention mechanism for effective fire detection," *Knowledge-Based Syst.*, vol. 329, Nov. 2025, doi: 10.1016/j.knosys.2025.114353.

[18] J. Simangunsong, M. S. Simanjuntak, and N. D. Simanjuntak, "Analisis Ketepatan Model CNN dalam Deteksi Asap Berbasis Citra," *J. Minfo Polgan*, vol. 14, pp. 2381–2390, Dec. 2025, doi: 10.33395/jmp.v14i2.15908.

[19] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J.*

Comput. Vis., vol. 115, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
[20] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[21] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*, International Machine Learning Society (IMLS), 2019, pp. 10691–10700. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html
[22] A. Howard *et al.*, "Searching for mobileNetV3," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, Jun. 2017, doi: 10.1145/3065386.
[24] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2014, pp. 3320–3328. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html
[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980
[26] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2017. [Online]. Available: https://arxiv.org/abs/1608.03983
[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.
[28] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," Nov. 2018, doi: 10.1109/WACV.2018.00097.
[29] P. P. Groumpos, "A Critical Historic Overview of Artificial Intelligence: Issues, Challenges, Opportunities, and Threats," *Artif. Intell. Appl.*, vol. 1, pp. 181–197, Jan. 2023, doi: 10.47852/bonviewAIA3202689.
[30] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in *ICLR 2021 - 9th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929
[31] R. Azad *et al.*, "Advances in medical image analysis with vision Transformers: A comprehensive review," Jan. 2024, *Elsevier B.V.* doi: 10.1016/j.media.2023.103000.