Stacking deep learning models for early detection of wildfire smoke plumes

Gonçalo Falcão
INESC-ID, Instituto Superior Técnico
Universidade de Lisboa
goncalo.a.falcao@tecnico.ulisboa.pt

Armando M. Fernandes
INOV & INESC-ID
armando.fernandes@inov.pt

Nuno Garcia
LASIGE, Faculdade de Ciências
Universidade de Lisboa
nrgarcia@ciencias.ulisboa.pt

Helena Aidos
LASIGE, Faculdade de Ciências
Universidade de Lisboa
haidos@ciencias.ulisboa.pt

Pedro Tomás
INESC-ID, Instituto Superior Técnico
Universidade de Lisboa
pedro.z.tomas@tecnico.ulisboa.pt

***Abstract*—Forest fires can have a devastating impact on the environment and pose a threat to human health. Due to the rapid rate at which they spread, early detection is critical to ensure a quick firefighting response. Automatic fire detection systems, based on machine learning, play a substantial role in this. Hence, in this paper, we combine the most recent state-of-the-art image classification models in order to produce a more robust and generalizable detection system through stacked generalization. In particular, we introduce a model stack composed of two types of model architectures: Convolutional Neural Networks and Vision Transformers. A meta-classifier, constituted by a small neural network, then learns how to best combine the predictions extracted by the models to identify smoke plumes. This approach exploits the architectural diversity and heterogeneity of the model stack to tackle the hardest-to-predict wildfire scenarios. Our results show an average accuracy of 96.5% and an Area Under the Precision-Recall Curve of 95.2%, which corresponds to an improvement of 2.11% and 1.2%, respectively, in comparison to the best-performing model from the stack.**

***Index Terms*—Forest fire detection, Image classification, Stacked generalization**

I. Introduction

The detection of early forest fires has become an increasingly relevant topic. Because of their fast spread rate, early detection allows firefighting teams to respond more effectively to potential fires, which can help minimize their damage. In their early stages, forest fires are characterized by small smoke plumes, easily confused with other occurrences, such as low clouds, fog, chimney smoke, and even dust caused by cars passing on nearby roads (see also Fig. 1). Hence, the problem lies in creating a detection system capable of accurately distinguishing real forest fires from any other kind of occurrence with similar visual behavior.

Recently, the use of Unmanned Aerial Vehicles (UAVs) and watch towers equipped with imaging sensors has become increasingly popular. On the other hand, the advancement of deep learning-based image classification models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) has made it possible to perform deeper and richer feature extraction that was previously much more limited [1]–[3]. Despite their increased performance, deep learning-based approaches still have difficulty detecting the hardest-to-predict scenarios since images can come in a diverse combination of weather and seasonal conditions. Smoke plumes consisting of early wildfires are also highly irregular, as they can appear in different sizes, shapes, textures, and intensities. Because of these diverse and irregular conditions, the models must be sufficiently robust to capture all types of fire signs, while minimizing the occurrence of false positives.

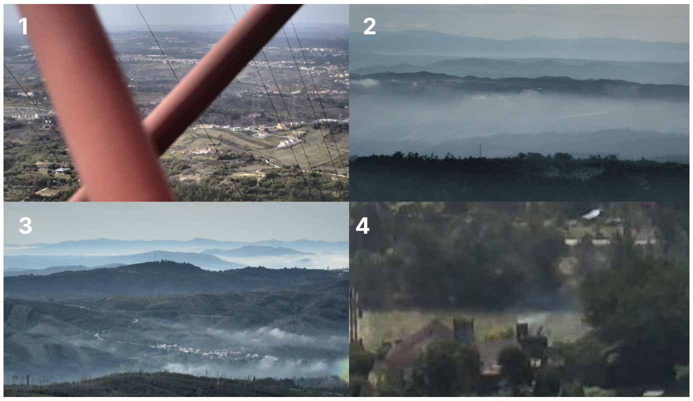

Fig. 1. Example of images taken from cameras mounted in surveillance towers with different types of occurrences hard to predict. 1: Image obstruction. 2: Foggy sky. 3: Low clouds. 4: Chimney smoke.

Inspired by the great performance of deep ensemble models in multiple areas such as speech, health care, and image classification [4], this paper introduces a model ensemble based on stacked generalization with the purpose of improving the performance of single-model detection. Stacked generalization, also known as stacked ensemble learning, involves training multiple base models on the same dataset and using

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PCIF/MPG/0051/2018 and through the research units INESC-ID (ref. UIDB/50021/2020) and LASIGE (ref. UIDB/00408/2020 and UIDP/00408/2020).

ISBN: 978-9-4645-9360-0

their predictions as inputs to a meta-classifier, which learns how to best combine those predictions and make a final decision. The meta-classifier can either be a simple pooling operation or a more complex neural network, as used in this work.

The chosen base models are constituted by two types of model architectures: CNN and ViT, more specifically, an EfficientNetV2 [5], a Data-Efficient Image Transformer (DeiT) [6], and a Swin Transformer [7]. The intuition behind this choice of models is to leverage the strengths of both architectures which have inherently different forms of learning [8]. Hence, by combining these two architecture types, we can potentially create a more robust model that can better handle both local and global features regarding fire signs in the images. This can be particularly helpful in detecting fire signs given the diversity and irregularity of early wildfires.

II. Related Work

A. Forest fire detection through deep learning

The problem of forest fire detection has been tackled by a wide variety of approaches based on different detection systems namely, terrestrial, aerial, and satellite-based [9]. CNNs are heavily used in the literature for both image classification and object detection. Xu *et al.* [10] proposed a mix of both through an ensemble learning method based on three learners with the aim of capturing more diverse features. The learners are made up of two object detectors: a YoloV5 and an EfficientDet that work in tandem to detect smoke and fire-like objects, and an EfficientNet that captures global information.

In [11] the authors propose a novel mosaic-based detection strategy that helps models to focus on the regions containing fire signs while taking to account the whole image. The authors used an EfficientNet as the detection model achieving an Area Under Receiver Operating Characteristic curve of 0.949 over the test set.

One problem with typical detection models is that the small size of smoke plumes associated with wildfires can make them difficult to detect, particularly when downsizing the images during training. This happens because these models receive as input considerably small image resolutions to constraint computational time. To tackle this, Perrolas *et al.* [12] propose a method based on the quad-tree search algorithm which consists of a recursive subdivision of space into four quadrants. In the context of fire detection, such method aims at recursively segmenting the images into smaller patches and individually searching for fire events in each. This way, the model can better focus on smaller regions of the image and capture smaller signs of smoke.

Other approaches to the detection of forest fires include the use of image segmentation models such as DeepLabV3+ and Fully Convolutional Networks (FCN) to detect segments of fire signs [13], [14]. The use of object detectors is also heavily seen especially with the use of the YOLO family of models given their speed and performance [15].

TABLE I
Model stack description

Model stack	Input resolution	# Parameters	Architecture
EfficientNetV2S	384x384	22M	CNN
DeiT	224x224	86M	ViT
Swin TransformerV2	224x224	87M	ViT

B. Ensemble learning for image classification

Ensemble learning is a popular technique for improving the performance of deep learning models. This approach has been shown to be effective in reducing overfitting and improving accuracy in a variety of image classification problems. Some of the most popular ensemble learning methods in deep learning are: Bagging, Boosting, and Stacking. Stacking is generally preferred when the model stack is heterogeneous while boosting and bagging are preferred for homogeneous model stacks [4].

Several forms of combining predictions include majority voting, weighted and unweighted mean, and logistic regression. All of these are well-studied and can produce different results depending on the task. Müller *et al.* [16] conducted a comprehensive study on the different ensemble methods by comparing their performances in medical image classification. The authors empirically demonstrated that stacking produced the best results, showing a performance gain of up to 13% in the F1-Score. Regarding the combination of the predictions, the authors got the best results with the simplest pooling functions, the top performing being mean and majority voting. They did not, however, try a neural network approach.

Ensemble learning approaches have been widely used in medical image classification because of their reliability and boost in performance [4]. Hameed *et al.* [17] proposed using a model ensemble for classifying carcinomas in breast cancer histopathology images. The authors employed an ensemble of fine-tuned VGG16 and VGG19 by averaging the extracted decision of both models, increasing performance. Xue et al. [18] propose a framework to classify cervical histopathological images through transfer learning and a weighted voting-based ensemble learning for combining the models' predictions. The authors show a substantial performance improvement of 2.5% to 5.5% accuracy higher than the base models.

III. Proposed Method

Inspired by the success of ensemble learning models in several fields of image classification, the main goal of this paper is to assess the use of model stacking to improve wildfire smoke plumes detection, by relying on a meta-classifier composed of three state-of-the-art classification models. Fig. 2 shows an overview of the proposed approach, as detailed in this section.

A. Model stack

The considered models to compose the proposed stacking method are based on CNNs and ViTs, which have substantial differences regarding their internal representation structure

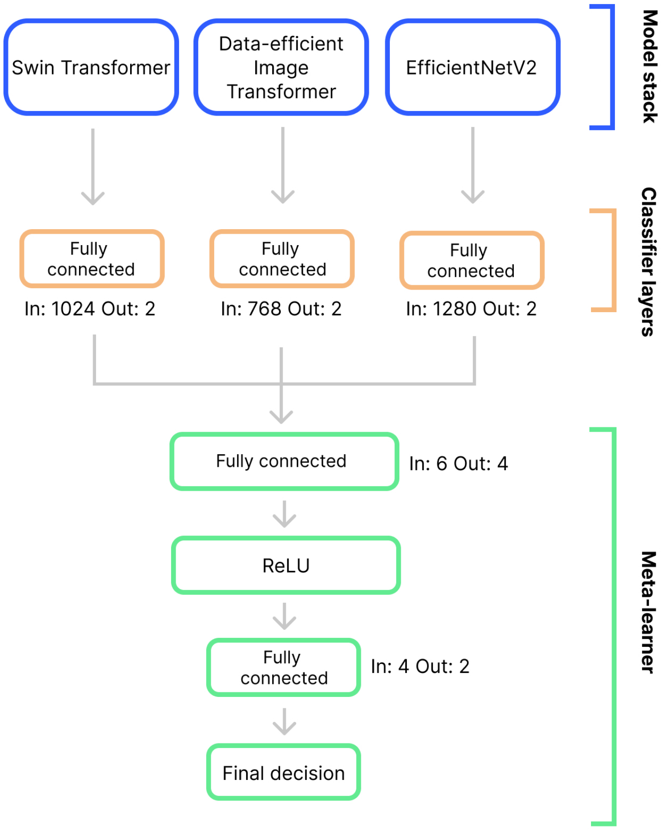

Fig. 2. Overview of the model stack and the architecture of the meta-learner. The diagram depicts the number of input parameters (In) and output (Out) for each fully connected layer.

[8]. ViTs, for instance, are capable of capturing more global information early in the lower layers because of the use of self-attention. They also have a strong ability to propagate that information throughout the layers, capturing long-range dependencies. CNNs on the other hand, focus more on local information at the lower layers, which is vital in the learning process [8], leading to measurably different features. These structural differences between the networks make this a heterogeneous stack which is preferred when using the stacking method.

The proposed stack comprises an EfficientNetV2, a DeiT, and a Swin transformerV2, as presented in Table I. For the EfficientNetV2, the small version (EfficientNetv2S) of the architecture, containing around 22 million parameters, was chosen to constrain inference time. It should be noticed that the medium-sized version of this network (EfficientNetv2M) would result in a doubling of the training and inference time for the same computational resources. As for the DeiT and the Swin Transformer V2, the base models contain around 86 million and 87 million parameters, respectively. The input images are downscaled to different resolutions for the two types of architectures. For the EfficientNetV2 model, the input image is downscaled to 384x384, whilst the other two models receive a 224x224 image.

B. *Meta-classifier architecture*

The key idea of stacking is to use the output of two or more models as input to another model, a meta-learner, which combines these intermediate predictions into a final prediction. Common choices for the meta-learner include linear regression, logistic regression, decision trees, random forests, gradient-boosting machines, and neural networks. It can also be a simple pooling function such as a mean or majority voting. In the case of a neural network, the meta-learner trains on the base model predictions and learns to weight them in the most optimal way. In this paper, we propose a simple neural network as our meta-learner, composed of two dense layers with a ReLU activation function in the middle (see Fig. 2). Compared to the traditional unweighted mean approach, the neural network showed better performance given its ability to learn more complex relations between the models' predictions.

C. *Training and evaluation approach*

All the models were independently pre-trained using the ImageNet dataset. Fine-tuning the models' parameters was attained using a small learning rate of 0.0001. Compared to just training the classifier layer and freezing all the layers below, fine-tuning the whole model produced around a 4 to 9% increase in accuracy. Overall, transfer learning was critical for boosting the models' performance.

Data augmentation techniques, such as RandAugment [19], also provided a significant performance boost. In particular, random horizontal flip, and color jitter were applied to reinforce the models' generalization ability. In addition, we normalized the images using the official mean and standard deviation of the ImageNet dataset. This step was necessary as we utilized pre-trained models that were trained on the ImageNet using such a normalization step.

The meta-learner was trained on the predictions given by the base models for every fold. Training on both subsets showed the best results compared to training on just one.

IV. Experimental Results

The experimental results were conducted in a Google Colab environment with the NVIDIA T4 Tensor Core GPU with 15GB of memory. Regarding the models, we utilized the pre-trained weights from the official PyTorch repository provided by the TorchVision library.

A. *Dataset*

The images used in this work were gathered from cameras mounted in nine surveillance towers installed at different geographical locations in the region of Leiria, Portugal, and located at distances from the sea between approximately 10 and 60 km. The cameras operate in the visible spectrum, forming a large-area wildfire surveillance system (see also Fig. 3). Given the characteristics of the surveillance system, the actual fire is not often visible in images (unless, for example, when it reaches high severity levels). Hence, for early forest fire identification, the system needs to be able to detect small smoke plumes and provide early alarms.

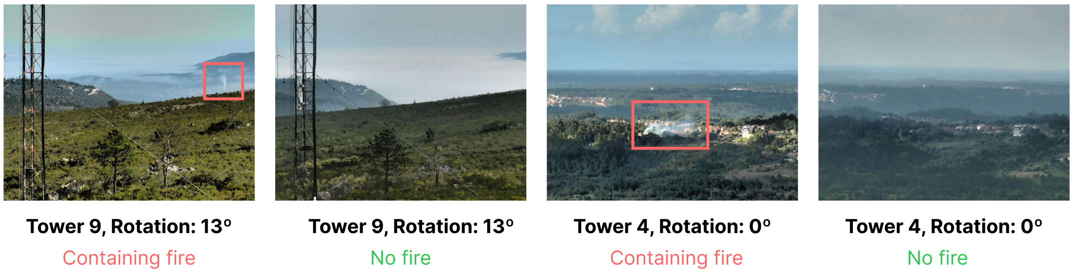

Fig. 3. Example of images taken from cameras mounted in surveillance towers. On the left, the tower 9 at a 13° rotation and on the right the tower 4 at a 0° rotation.

On total, the currently used dataset contains 14135 images with smoke plumes and 21205 without smoke plumes, which results in a 40%-60% class imbalance. This data imbalance is however more pronounced in particular towers, as observed in Table II. In general, the number of images without smoke plumes exceeds those with them, with the greatest difference being observed in tower 1. Specifically, only 10% of the images captured in tower 1 contain a smoke plume, while a vast majority of 90% depict no fire activity.

TABLE II
NUMBER OF IMAGES IN BOTH CLASSES (FIRE AND NO FIRE) FOR EACH OF THE 9 TOWERS (1,2,3,4,7,8,9,10,11)

#	1	2	3	4	7	8	9	10	11
Fire	206	2796	475	929	2070	3571	2351	1130	607
No fire	1944	2920	3315	2393	1304	2329	2409	2822	1769

B. Quantitative results

To train and effectively evaluate our ensemble model, we applied a leave-one-out technique for each tower for evaluation, while training on the remaining. This guaranteed that model testing was performed in a very different environment from the one it was trained in since the towers are placed in distinct geographical locations. Given the time and computational restrictions for evaluation, only three folds (or towers) will be evaluated, corresponding to towers 4, 8, and 10, making it a 3-fold cross-validation.

To assess the individual models' performance and compare them against the stacking approach, several metrics were used, namely Accuracy (class-weighted, to account for class imbalance), Precision, Recall, F1-Score and Area Under the Precision-Recall Curve (AUPRC).

Table III shows the results of each base model and the ensemble method for 3 sample folds corresponding to the leave-one-out of towers 4, 8, and 10. The proposed neural network meta-classifier presents the overall best results in every fold compared to all of the base models in the stack. Noticeably, when comparing with the best base model, there is a substantial increase in the AUPRC score by a mean of 1.20%, and in the accuracy by a mean of 2.11% (representing a $\approx$37% average reduction of miss-classified images). Table IV shows the average and standard deviation values for each metric. Overall, the proposed stacking approach has higher values in all the performance metrics, except for the precision which has a similar average performance as the EfficientNetV2. Also, the standard deviation is smaller for the model stack, illustrating the robustness of the model compared to a single model.

Fig. 4 shows some examples of true positives and false negatives predicted by the proposed stacking approach. It is particularly interesting to observe that in the top right image, the model struggles with the haze, not capturing the small smoke plume in the distance.

It is also interesting to compare the proposed approach with the work of Fernandes *et al* [11], which used the same dataset. In their work, the authors attained the best results using an EfficientNet (v1) model. Here, we propose a model stacking approach which shows to provide better results than a single updated EfficientNet (v2).

While model stacking seems to provide improved results over the base model, it is also important to analyze the execution time, particularly to validate whether such a system can be used in real-time. The proposed model stacking approach takes a mean of 58ms to provide the results over a batch of 16 images (which compares with an average of 17ms for the base models). This provides the ability to classify up to 275 towers, considering an average frame rate of 1 image per second (which is a practical value given the natural speed and spread of forest fires) - currently, the image capture frame rate is below 1 image per minute. This validates the use of the proposed model stacking approach in real-time scenarios.

V. Conclusion

In this paper, we explored the potential of ensemble methods for the task of early forest fire detection. We introduced an ensemble stacking model that effectively improves the performance of the base models by 1.3% to 2.2% in accuracy. Our ensemble contains a meta-classifier composed of a small neural network that learns how to best combine the predictions from the model stack. The stack is composed of state-of-the-art image classification architectures which given their different forms of interpreting visual information, together cooperate for enhanced accuracy and robustness in recognizing fire signs. For evaluating our approach, we used a leave-one-tower-out cross-validation method, reaching an average accuracy of 96.46% and Area Under the Precision-Recall Curve of 95.2%.

TABLE III
PERFORMANCE EVALUATION OF THE BASE MODELS AND THE PROPOSED ENSEMBLE STACK WITH A META-CLASSIFIER.

	Fold 4					Fold 8					Fold 10				
Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUPRC (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUPRC (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUPRC (%)
EfficientNetV2	95.33	95.18	96.19	94.73	93.84	91.78	89.87	**98.53**	83.69	90.19	95.93	97.04	99.07	95.36	97.84
DeiT	88.77	89.18	95.52	84.65	89.85	91.02	88.30	94.10	84.62	86.50	93.72	95.50	97.33	94.13	95.92
Swin TransformerV2	95.15	94.25	93.95	96.42	93.93	93.15	90.75	94.60	88.43	89.08	94.63	96.16	**99.24**	**99.22**	97.69
Proposed	**96.75**	**96.73**	**97.48**	**96.43**	**95.84**	**95.36**	**93.15**	96.23	**91.32**	**91.53**	**97.27**	**97.95**	98.85	97.24	**98.06**

TABLE IV
AVERAGE ± STANDARD DEVIATION OF THE MEASURED METRICS FOR THE BASE MODELS AND THE PROPOSED ENSEMBLE STACK APPROACH.

	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUPRC (%)
EfficientNetV2	95.33 ±2.09	95.18 ±3.60	**96.19** ±0.57	94.73 ±5.90	93.84 ±3.85
DeiT	91.17 ±1.52	90.99 ±3.64	95.65 ±1.62	87.80 ±4.84	90.76 ±4.72
Swin TransformerV2	94.31 ±0.78	93.72 ±2.71	95.93 ±2.39	94.69 ±5.42	93.57 ±4.31
Proposed	**96.46** ±0.96	**95.94** ±2.41	97.52 ±1.31	**95.00** ±2.99	**95.14** ±3.27

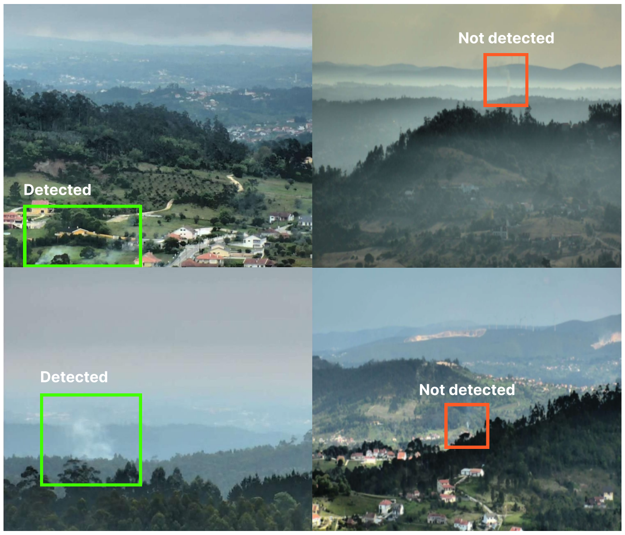

Fig. 4. Examples of predictions by the ensemble model. Left: true positives (green bounding box); right: false negatives (red bounding box). The bounding boxes are simply illustrative to know the position of the fire sign in the image.

REFERENCES

[1] Y. Luo, L. Zhao, P. Liu, and D. Huang, "Fire smoke detection algorithm based on motion characteristic and convolutional neural networks," *Multimedia Tools and Applications*, vol. 77, pp. 15 075–15 092, 2018.

[2] Q.-x. Zhang, G.-h. Lin, Y.-m. Zhang, G. Xu, and J.-j. Wang, "Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images," *Procedia engineering*, vol. 211, pp. 441–446, 2018.

[3] S. Khan, K. Muhammad, T. Hussain, J. Del Ser, F. Cuzzolin, S. Bhattacharyya, Z. Akhtar, and V. H. C. de Albuquerque, "DeepSmoke: Deep learning model for smoke detection and segmentation in outdoor environments," *Expert Systems with Applications*, vol. 182, 2021.

[4] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.

[5] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[6] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[8] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[9] P. Barmpoutis, P. Papaioannou, K. Dimitropoulos, and N. Grammalidis, "A review on early forest fire detection systems using optical remote sensing," *Sensors*, vol. 20, no. 22, p. 6442, 2020.

[10] R. Xu, H. Lin, K. Lu, L. Cao, and Y. Liu, "A forest fire detection system based on ensemble learning," *Forests*, vol. 12, no. 2, p. 217, 2021.

[11] A. M. Fernandes, A. B. Utkin, and P. Chaves, "Automatic early detection of wildfire smoke with visible light cameras using deep learning and visual explanation," *IEEE Access*, vol. 10, pp. 12 814–12 828, 2022.

[12] G. Perrolas, M. Niknejad, R. Ribeiro, and A. Bernardino, "Scalable fire and smoke segmentation from aerial images using convolutional neural networks and quad-tree search," *Sensors*, vol. 22, no. 5, p. 1701, 2022.

[13] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 04, pp. 640–651, 2017.

[14] F. Yuan, L. Zhang, X. Xia, B. Wan, Q. Huang, and X. Li, "Deep smoke segmentation," *Neurocomputing*, vol. 357, pp. 248–260, 2019.

[15] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[16] D. Müller, I. Soto-Rey, and F. Kramer, "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks," *Ieee Access*, vol. 10, pp. 66 467–66 480, 2022.

[17] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. Maria Vanegas, "Breast cancer histopathology image classification using an ensemble of deep learning models," *Sensors*, vol. 20, no. 16, p. 4373, 2020.

[18] D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104 603–104 618, 2020.

[19] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.