# *Improving Deep Learning-Based Wildfire Smoke Plume Detection with a Multi-Model Ensemble Approach*

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

With increasing frequency and severity of wildfires, there is an urgent need for wildfire and smoke detection tools that can effectively and rapidly monitor smoke at a large scale. Recent advancements in computer vision have demonstrated the potential of deep learning models to automatically separate high-resolution images into labeled regions for high-accuracy feature detection. However, single-model approaches can struggle with generalization and accuracy in diverse conditions, which is necessary in the operational use of a smoke detection tool. To address these challenges, we propose using an ensemble of deep learning models to produce more accurate annotations of wildfire smoke plumes and their relative density (light, medium, heavy) in satellite imagery. Our preliminary results indicate that ensemble techniques can improve performance compared to using a single model, and that further investigation is needed to optimize the ensemble. This approach aims to provide a more reliable satellite-based tool for real-time monitoring of smoke. This will have numerous downstream impacts, such as aiding fire and hazard management efforts, improving modeling of wildfire behavior and air quality, and ultimately contributing to climate resilience and adaptation strategies.

## 1 Introduction

In the last four decades, wildfire activity has increased drastically in the U.S. In fact, the conditions leading to wildfires have been shown to occur more frequently as a direct result of climate change [1]. Satellite observations reveal that the number of days with smoke in the air have substantially increased in the U.S. during the last two decades [1]. Furthermore, the human impacts of smoke exposure include increased morbidity and mortality as well as downstream economic costs [2]. Therefore, it is essential to develop effective, large-scale smoke monitoring tools.

Satellite imagery can be used to detect and monitor the evolution of smoke over large areas. However, such methods have yet to provide precise and high-frequency information on smoke density, or are confined to small case-study regions [3–5].

## 2 Background

The National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environmental Satellites (GOES) provide high spatial and temporal resolution imagery of North America [6]. The NOAA Hazard Mapping System (HMS) Fire and Smoke Product currently relies on expert human analysts to annotate the presence of smoke over North America using GOES imagery [7].

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

However, this product is limited by the availability of human analysts and their time, with annotations outputted only once to a few times a day and usually having a delay between smoke occurence and the annotation. Thus, emergency responders may be delayed without real-time smoke conditions, or could miss early detections of fire considering that smoke can obfuscate fire visibility. To address these limitations, we leverage advancements in DL to automate the detection of smoke from GOES imagery, using the existing HMS dataset for training. By automating this task, we can enable more frequent detection of smoke plumes, which will aid in active wildfire monitoring and mitigating air quality impacts.

Ensembles, which combine the predictions of multiple models, have been shown to often perform better than a single model in classification tasks [8, 9]. Furthermore, combining the predictions of multiple independently-trained neural networks can improve generalization and accuracy [10–12]. [13] demonstrated how a stacked ensemble of deep learning (DL) models can be used improve classification of smoke plumes in surveillance camera imagery in Leiria, Portugal. To increase the scale at which smoke can be monitored, we propose using an ensemble of DL models for detecting smoke with large-scale satellite imagery. To our knowledge, these ensemble methods have not been applied to smoke detection in GOES imagery at the continental scale of North America.

# 3 Proposed methods

We use the SmokeViz dataset (Section 7.1) which consists of 183,672 smoke plume samples spanning 2018-2024, each with three spectral channels of GOES imagery paired with human analyst annotations of light, medium, or heavy smoke. We use 2018-2021 and 2024 of training, 2023 for validation, and 2022 for the test set, ensuring the testing and validation data years are independent of the training data. During validation and testing, we quantify model performance with Intersection over Union (IoU) score, which measures pixel-wise alignment between the model prediction and the ground truth. The IoU metric supports pathways to climate impact since improved IoU score directly relates to more smoke being accurately detected (reducing false detections and increasing true detections) which equips hazard management to appropriately respond to current wildfire and smoke conditions.

We utilize a variety of encoder-decoder architectures designed for semantic segmentation that include different features such as multi-scale fields-of-view and precise boundary detection [14–16]. Additionally, we selected the best-performing single architecture and trained it with 12 seeds to generate different initial random weights. By varying the initialization, the model's parameter space can be searched more fully; ideally, different minima of the loss function will be found. All the models were trained independently for 24 hours on 8 Nvidia P100 GPUs using the Adam optimizer and the binary cross entropy loss function.

The ensemble method we are using in this preliminary analysis is an unweighted average of $N$ model outputs (Figure 1), as used in [17, 11]. We experimented with ensemble size (supplementary section 7.2), and present results here with $N = 8$. We create an architecture-based ensemble, comprised of the architectures and encoders with the best individual performance, and a initialization-based ensemble, comprised of models with the same architecture (PAN) but different random initializations.

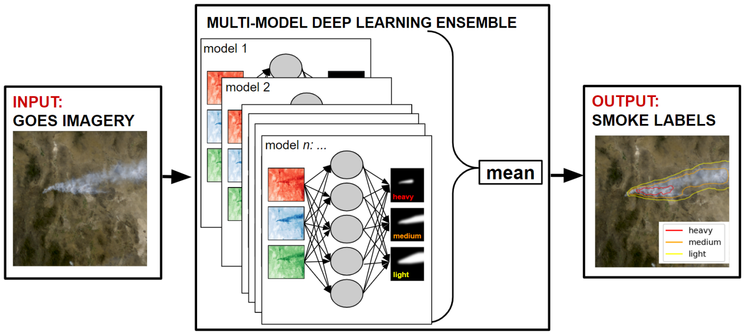

Figure 1: Multi-model ensemble framework. GOES imagery is inputted to $N$ independently-trained models whose output is combined with an unweighted average to produce the ensemble prediction of pixel-wise smoke labels.

## 4 Preliminary results

Table 1 shows the IoU scores on the test set for individual models and ensembles. We report the 2 best-performing architectures, DLV3P [14] and PAN [15], with additional results in supplementary section 7.3. Both the architecture ensemble and the initialization ensemble outperform the individual models across all IoU metrics. Considering that the test set is independent from the training and validation set, this improvement in performance can be interpreted as an improvement in model generalization. Further, Figure 2 shows an example of smoke plume detection from the testing dataset. The ensemble predictions have smoother boundaries than the individual model outputs, making the prediction more comparable to the human analyst-drawn annotations. Future work will investigate the mechanisms behind the ensemble's improvement in performance and smoothing of boundaries, as well as strategies for optimal ensemble selection.

Table 1: Test set IoU results across heavy, medium and light smoke density and over all densities with single models and ensemble schemes.

| | Heavy | Medium | Light | Overall |
|---|---|---|---|---|
| Single Model: DLV3P [14] | 0.347 | 0.441 | 0.666 | 0.599 |
| Single Model: PAN [15] | 0.349 | 0.478 | 0.664 | 0.604 |
| Architecture Ensemble | 0.400 | 0.507 | 0.692 | 0.635 |
| Initialization Ensemble | 0.409 | 0.512 | 0.684 | 0.631 |

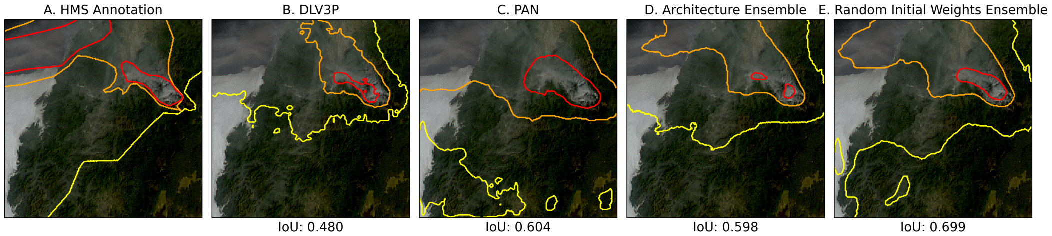

Figure 2: Example of smoke plume detection at (43.37°, -123.25°) on 2022/10/15 15:50 UTC. Red, orange, and yellow contours represent heavy, medium and light density smoke annotation/prediction, respectively. (A) displays the ground truth annotation; (B-C) show the predictions of two individual models; (D) shows the prediction of the architecture-based ensemble; (E) shows the prediction of an ensemble made with models initialized with different random weights.

## 5 Limitations and future work

This proposal explores two schemes for building ensembles of DL models that both improve on testing set IoU and smooth annotation boundaries. However, further investigation is required to give insight on how the ensemble reduces error and improves generalizability, as well as what the optimal ensemble size and type are. One area to explore is model stacking, where an optimized meta-model is used to combine multi-model outputs, as used in [13, 17–20]. Furthermore, future work will utilize the multi-model ensemble to quantify uncertainty in smoke annotations, enabling users like wildfire response teams and environmental agencies to assess the reliability of detections in real time.

## 6 Pathways to climate impact

We propose using multi-model ensembles for smoke detection from satellite imagery and show preliminary results indicating that a DL ensemble can improve model performance. These techniques can be used to aid fire and hazard management by automating the monitoring of smoke in real-time from large-scale satellite imagery with smooth and accurate smoke annotations. This will improve prediction of wildfire movement and air quality impacts, ultimately supporting climate resilience and adaptation strategies.

# References

[1] Marshall Burke, Anne Driscoll, Sam Heft-Neal, Jiani Xue, Jennifer Burney, and Michael Wara. The changing risk and burden of wildfire in the united states. *Proceedings of the National Academy of Sciences*, 118(2):e2011048118, 2021.

[2] Wayne E. Cascio. Wildland fire smoke and human health. *The Science of the Total Environment*, 624:586–595, 05 2018. Epub 2017 Dec 27.

[3] Jeff Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery. *ArXiv*, abs/2109.01637, 2021.

[4] Jiayun Yao, Sean M. Raffuse, Michael Brauer, Grant J. Williamson, David M.J.S. Bowman, Fay H. Johnston, and Sarah B. Henderson. Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the calipso satellite. *Remote Sensing of Environment*, 206:98–106, 2018.

[5] Alexandra Larsen, Ivan Hanigan, Brian J. Reich, Yi Qin, Martin Cope, Geoffrey Morgan, and Ana G. Rappold. A deep learning approach to identify smoke plumes in satellite imagery in near-real time for health risk communication. *Journal of Exposure Science & Environmental Epidemiology*, 31(1):170–176, 2021.

[6] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R Series: A New Generation of Geostationary Environmental Satellites*. Elsevier, 2019.

[7] Donna McNamara, George Stephens, Mark Ruminski, and Tim Kasheta. The hazard mapping system (hms) - noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference on Satellite Meteorology and Oceanography*, 01 2004.

[8] Thomas G. Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15, 2000.

[9] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[10] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[11] Aurélien Bibaut Cheng Ju and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018. PMID: 31631918.

[12] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001.

[13] Gonçalo Falcão, Armando M. Fernandes, Nuno Garcia, Helena Aidos, and Pedro Tomás. Stacking deep learning models for early detection of wildfire smoke plumes. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1370–1374, 2023.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[15] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *CoRR*, abs/1805.10180, 2018.

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

[17] Manthena Sivanuja, P.J.R Shalem Raju, M. Prasad, Raja Rao PBV, K. Satish Kumar, and P. Kiran Sree. A novel ensemble-based deep learning framework combining cnn and transfer learning models for enhanced wildfire detection. In *2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, pages 1–6, 2025.

[18] Linh Nguyen Van and Giha Lee. Optimizing stacked ensemble machine learning models for accurate wildfire severity mapping. *Remote Sensing*, 17(5), 2025.

[19] Binxu Zhai and Jianguo Chen. Development of a stacked ensemble model for forecasting and analyzing daily average pm2.5 concentrations in beijing, china. *Science of The Total Environment*, 635:644–658, 2018.

[20] Jiayue Gu, Shuguang Liu, Zhengzheng Zhou, Sergey R. Chalov, and Qi Zhuang. A stacking ensemble learning model for monthly rainfall prediction in the taihu basin, china. *Water*, 14(3), 2022.

# 7 Supplementary material

## 7.1 Data and code availability

The code for this work is available at `https://github.com/anonymous-ensemble-smoke/ensemble-AI-smoke-detection/tree/main`. The dataset can be accessed at `https://noaa-gsl-experimental-pds.s3.amazonaws.com/index.html#SmokeViz/`.

## 7.2 Ensemble size analysis

Figure 3 shows the IoU performance over all smoke densities as a function of ensemble size, $N$, for the two ensemble schemes. The ensemble with different initial weights generally improves as models are added to the ensemble. The ensemble of different architectures improves with more models up to 8 models, but then decreased in IoU with more models added to the ensemble. This decrease in performance could be due to the additional architectures not having enough variation in model bias to improve ensemble performance. Future work will aim to clarify exactly how different ensemble sizes behave and reduce error.

An additional example from the test data set is shown in Figure 4, where the individual model output has jagged boundaries and the ensemble outputs smooth over these edges. We see a peak in performance at $N = 8$ in this sample where the $N = 8$ ensemble has the highest IoU score, and the smoothing does not seem to improve in the $N = 12$ ensemble output. This sample supports the proposed idea that ensemble DL can smooth over rough edges in semantic segmentation, and warrants further investigation for the optimal ensemble and how to use the multi-model approach to quantify uncertainty.

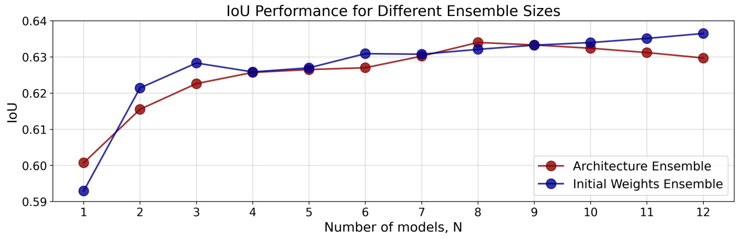

Figure 3: Overall IoU as a function of $N$ for two ensemble design schemes: random initial weights (blue) and architecure-based (red).

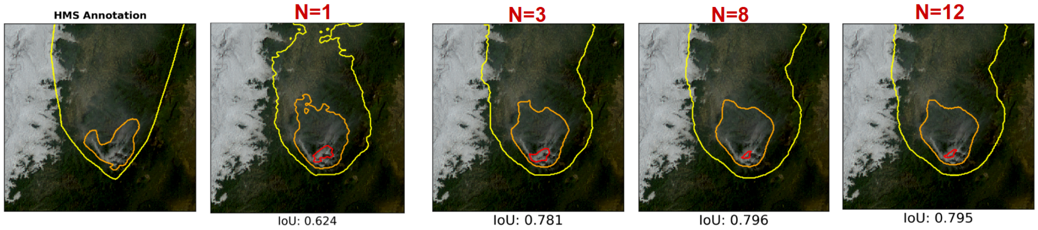

Figure 4: Example of smoke plume detection at (44.24, -122.74) on 2022/09/27 15:30 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. The first panel displays the ground truth HMS annotation; the second panel is the individual model output of DLV3P; the following panels the prediction of an architecture-based ensemble as it increases in size, $N$.

### 7.3 Architecture analysis

We assessed the performance of multiple architectures on our test set (Table 2). We also experimented with different sizes in the encoder backbone, EfficientNet, where b1 is the smallest encoder, and b3 is the largest encoder we used.

Table 2: Test set IoU results across heavy, medium and light smoke density and over all densities using numerous architectures.

| Architecture | Encoder | Heavy | Medium | Light | Overall |
|---|---|---|---|---|---|
| PAN | efficientnet-b2 | 0.349 | 0.478 | 0.664 | 0.604 |
| PAN | efficientnet-b1 | 0.364 | 0.468 | 0.648 | 0.590 |
| DLV3P | efficientnet-b2 | 0.347 | 0.441 | 0.666 | 0.599 |
| DLV3P | efficientnet-b3 | 0.365 | 0.474 | 0.653 | 0.595 |
| Unet++ | efficientnet-b1 | 0.369 | 0.472 | 0.654 | 0.598 |
| Unet++ | efficientnet-b2 | 0.354 | 0.464 | 0.662 | 0.597 |
| PSPNet | efficientnet-b2 | 0.374 | 0.482 | 0.651 | 0.596 |
| MANet | efficientnet-b2 | 0.352 | 0.478 | 0.646 | 0.587 |
| LinkNet | efficientnet-b2 | 0.360 | 0.470 | 0.621 | 0.570 |

### 7.4 Intersection Over Union (IoU) formula

Equation 1 provides a mathematical formula for IoU, where $y_i$ represents the ground truth and $y_i^*$ represents the model's prediction.

$$\text{IoU}_{\text{overall}} = \sum_{i=\text{light}}^{\text{heavy}} |y_i \cap y_i^*| \div \sum_{i=\text{light}}^{\text{heavy}} |y_i| \cup |y_i^*| \tag{1}$$