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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

1 Introduction

18 In the last four decades, wildfire activity has increased drastically in the U.S. In fact, the conditions
19 leading to wildfires have been shown to occur more frequently as a direct result of climate change [1].
20 Wildfires produce particulate emissions, such as smoke and ash, which can be tracked to quantify
21 the impact a fire has beyond its burn area. Satellite observations reveal that the number of days
22 with smoke above the continental U.S. have substantially increased in the last two decades [1].
23 Furthermore, the human impacts of smoke exposure include increased morbidity and mortality as well
24 as downstream economic costs [2].

25 Determining causal links between wildfire activity and health impacts in regions distant from the
26 source fire requires accurate large-scale monitoring of wildfire smoke and its movement. An intuitive
27 approach is to utilize satellite-based methods that can monitor evolution of smoke plumes over large
28 areas. However, such methods have yet to provide precise and high-frequency information on smoke
29 density, or are confined to small case-study regions [3–5].

30 2 Background

31 2.1 Current large-scale, satellite-based smoke detection

32 The National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environ-
33 mental Satellites (GOES) provide high spatial and temporal resolution imagery of North America
34 [6]. The NOAA Hazard Mapping System (HMS) Fire and Smoke Product currently relies on expert
35 human analysts to annotate the presence of smoke over North America using GOES imagery [7].
36 However, this product is limited by the availability of human analysts and their time, with annotations
37 outputted only once to a few times a day and usually having a delay between smoke occurrence and
38 the annotation. Thus, emergency responders may be delayed without real-time smoke conditions, or
39 could miss early detections of fire considering that smoke can obfuscate fire visibility. To address
40 these limitations, we leverage advancements in deep learning to automate the detection of smoke
41 from GOES imagery, using the existing HMS dataset for training. By automating this task, we can
42 enable more frequent detection of smoke plumes, which will aid in active wildfire monitoring and
43 mitigating air quality impacts.

44 2.2 Ensemble techniques

45 This proposal focuses on enhancing the capability of deep learning models to detect smoke using
46 multi-model ensemble methods. Ensembles, which combine the predictions of multiple models,
47 have been shown to often perform better than a single model in classification tasks [8]. Particularly,
48 utilizing a diverse set of classifiers in an ensemble is important to achieve the improvement in
49 performance [9]. Furthermore, combining the predictions of multiple independently-trained neural
50 networks can improve generalization and detection accuracy [10–12]. To our knowledge, these
51 ensemble methods have not yet been used for wildfire smoke detection.

52 3 Methods

53 We use the SmokeViz dataset which consists of 183,672 smoke plume samples, each with three
54 spectral channels (C01-C03) of GOES imagery paired with human analyst smoke annotations (pixel-
55 wise labels of smoke density of light, medium, or heavy). The data spans 2018-2024, and we use
56 2023 for the validation set and 2022 for the test set, with the remaining years used for training. This
57 ensures the testing and validation data are independent of the training data. During validation and
58 testing, we quantify model performance with Intersection over Union (IoU) score, which measures
59 pixel-wise alignment between the model prediction and the ground truth. The IoU metric supports
60 pathways to climate impact since improved IoU score directly relates to more smoke being accurately
61 detected (reducing false detections and increasing true detections) which equips hazard management
62 to appropriately respond to current wildfire and smoke conditions.

63 We utilize a variety of encoder-decoder architectures designed for semantic segmentation that include
64 different features such as multi-scale fields-of-view and precise boundary detection [13–15]. Addi-
65 tionally, we selected the best-performing single architecture and trained it with 12 seeds to generate
66 different initial random weights. By varying the initialization, the model’s parameter space can be
67 searched more fully; ideally, different minima of the loss function will be found. All the models were
68 trained independently for 24 hours on 8 Nvidia P100 GPUs using the Adam optimizer and the binary
69 cross entropy loss function.

70 The ensemble method we are using in this preliminary analysis is an unweighted average of N
71 model outputs (Figure 1), as used in [11]. We experimented with the ensemble size in supplementary
72 section 7.2, and present results here with $N = 8$. We create an architecture-based ensemble and a
73 initialization-based ensemble.

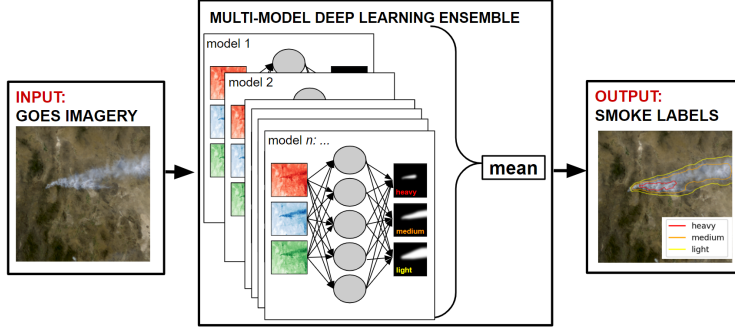


Figure 1: Multi-model ensemble framework. GOES imagery is inputted to N independently-trained models whose output is combined with an un-weighted average to produce the ensemble prediction of pixel-wise smoke labels.

4 Preliminary results

Table 1 shows the IoU scores on the test set for individual models and ensembles. The ensemble of 8 different architectures outperforms the individual models, with an improvement in all IoU metrics. The ensemble of 8 different initial weights (but the same architecture, PAN) also outperforms the individual models, with a similar improvement in the IoU scores. This improvement is likely due to the different initializations leading to the models searching different parts of the parameter space and thus finding different minima of the loss function. Future work is necessary to reveal the mechanisms behind the ensemble’s improvement in performance, as well as how to optimally select models that are included in the ensemble.

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.

Table 1: Test 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 [13]	0.347	0.441	0.666	0.599
Single Model: PAN [14]	0.349	0.478	0.664	0.604
Architecture Ensemble	0.400	0.507	0.692	0.635
Random Initial Weights Ensemble	0.409	0.512	0.684	0.631

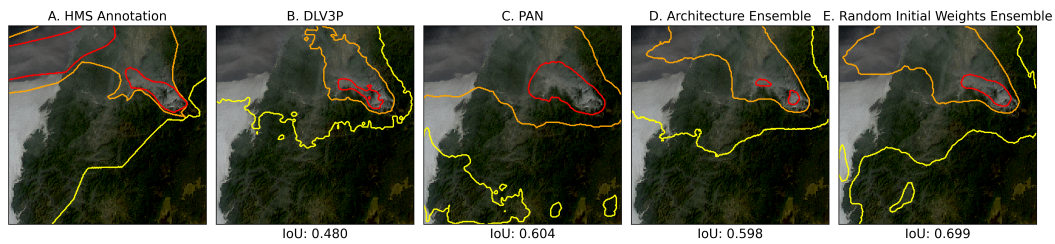


Figure 2: Example of smoke plume detection at $(43.37^\circ, -123.25^\circ)$ on 2022/10/15 15:50 UTC. Red, orange, and yellow contours represent heavy, medium and light density smoke annotation/prediction, respectively. (A) displays the ground truth annotation; (B-C) show the predictions of two individual 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 deep learning models that both improve on testing set IoU and smooth annotation boundaries. However, further investigation is required to

89 give insight on how the ensemble reduces error and improves generalizability and what the optimal
90 ensemble size and type are. One area to explore is "stacking", where an optimized meta-model is
91 used to combine multi-model outputs [16, 17]. Furthermore, future work will utilize the multi-model
92 ensemble to quantify uncertainty in smoke annotations, enabling users like wildfire response teams
93 and environmental agencies to assess the reliability of detections in real time.

94 **6 Pathways to climate impact**

95 These ensemble techniques can be used to aid fire and hazard management by automating the
96 monitoring of smoke in real-time from satellite imagery with smooth and accurate smoke annotations.
97 This will improve prediction of wildfire movement and air quality impacts, ultimately supporting
98 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] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.
- [14] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *CoRR*, abs/1805.10180, 2018.
- [15] 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.
- [16] 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.
- [17] 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 deep learning 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.

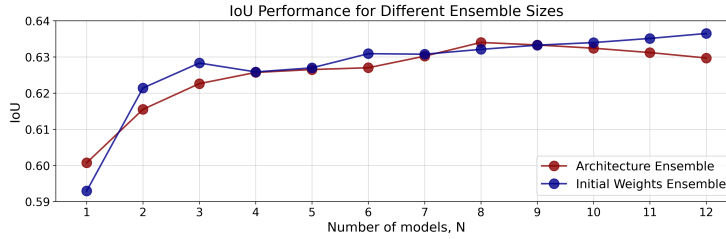


Figure 3: Overall IoU as a function of N for two ensemble design schemes: random initial weights (blue) and architecture-based (red).

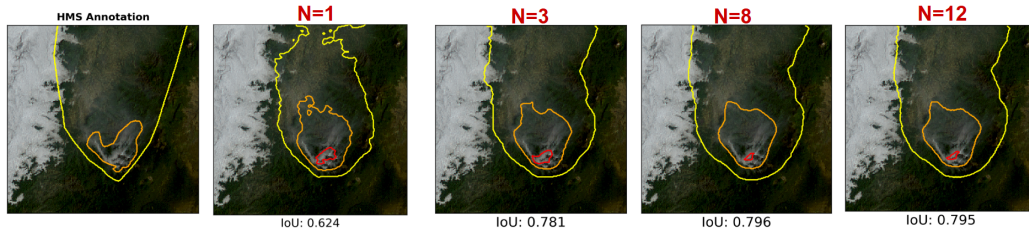


Figure 4: Example of smoke plume detection at (44.24, -122.74) on 2022/09/27 15:30 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. The first panel displays the ground truth HMS annotation; 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 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.

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