IMPROVING SATELLITE-BASED WILDFIRE SMOKE PLUME DETECTION WITH DEEP ENSEMBLES

Annabel Wade[1], Rey Koki[2,3,4], Christina Kumler-Bonfanti[2,3,4], Dale Durran[5]
[1]Boston University, [2]NOAA Global Systems Laboratory,
[3]Cooperative Institute for Research in Environmental Sciences,
[4]University of Colorado Boulder, [5]University of Washington
wade@bu.edu

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 machine learning to automatically label regions of high-resolution images with high accuracy. However, single-model approaches can struggle with generalization and accuracy in diverse conditions, which is necessary for operational smoke detection. 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 results indicate that deep ensemble techniques improve performance compared to using a single model. This approach aims to provide a more reliable satellite-based tool for real-time smoke monitoring, thereby aiding fire and hazard management efforts and improving the modeling of wildfire behavior and air quality.

1 BACKGROUND

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]. Further, smoke exposure has been associated with increased morbidity and mortality, as well as downstream economic costs [2]. Thus, 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, current methods have yet to provide precise and high-frequency information on smoke density, or are confined to small case-study regions [3; 4; 5].

Deep ensembles, which combine the predictions of multiple independently-trained neural networks, can improve model generalization and accuracy compared to single models [6; 7; 8]. For instance, Falcão et al. demonstrated using a deep ensemble to improve classification of smoke plumes in local surveillance imagery [9]. To increase the scale at which smoke can be monitored, we propose using a deep ensemble for detecting smoke with continental-scale satellite imagery. These methods are applied to North America, but could be applied to other regions where wildfire smoke detection is important, given that sufficient observations and matching labels are available.

The National Oceanic and Atmospheric Administration (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 Geostationary Operational Environmental Satellites (GOES) imagery [10; 11]. However, this product is limited by the availability of human analysts and their time. These annotations are outputted only once to a few times per day and usually have a delay between smoke occurence and the reported detection. Thus, emergency response to wildfires may be delayed without real-time smoke conditions and without early fire detection when smoke obfuscates fire visibility. To address these limitations, we leverage advancements in deep learning to automate the detection of smoke from GOES imagery, which will ultimately 1) enable more frequent detection of smoke plumes, 2) aid in active wildfire monitoring, and 3) mitigate air quality impacts.

2 METHODS

2.1 DATASET AND EVALUATION

We use the SmokeViz dataset [12] which consists of 183,672 smoke plume samples, each with three spectral channels of GOES imagery (C01-C03) paired with human analyst annotations of light, medium, or heavy smoke. We use 2018-2021 and 2024 for training, 2023 for validation, and 2022 for the test set, ensuring the testing and validation data years are independent of the training data. SmokeViz has samples until November 2024, so we chose 2022 and 2023 as the validation and test years to capture a full year of wildfire activity in both the validation and test sets. 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 (equations in supplemental section A.3). Improving the IoU score directly relates to more smoke being accurately detected, reducing false detections and increasing true detections.

2.2 MODEL ARCHITECTURES AND TRAINING

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 [13; 14; 15]. Additionally, we selected the best-performing single architecture and trained it with 12 seeds to generate different initial random weights to find different minima of the loss function. All the models were trained independently on 8 Nvidia P100 GPUs using the Adam optimizer. The binary cross entropy loss function is used with thermometer encoding (no smoke $= [0\ 0\ 0]$, light smoke $= [0\ 0\ 1]$, medium smoke $= [0\ 1\ 1]$, heavy smoke $= [1\ 1\ 1]$) to account for the ordinal nature of the smoke density labels.

2.3 ENSEMBLE METHODOLOGY

The ensemble method we are using in this analysis is an unweighted average of N model outputs (Figure 1), as used in [16; 7]. We present results here with $N = 8$ since it was found to be an optimal ensemble size based on empirical resutls (see supplemental section A.2). We create an **architecture-based ensemble**, comprised of the architectures and encoders with the best individual test set performance, and an **initialization-based ensemble**, comprised of models with the same architecture (PAN [14]) but different random initializations.

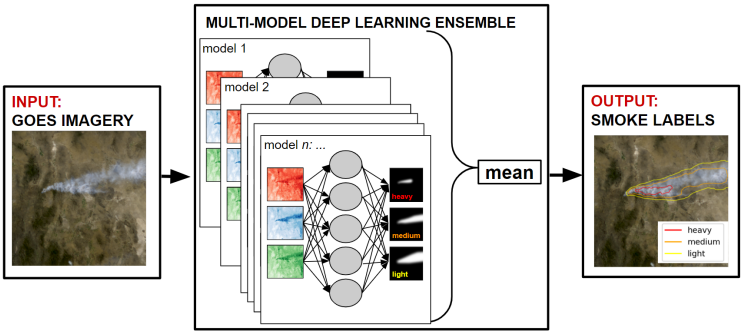

Figure 1: Multi-model ensemble framework. GOES imagery is inputted to N independently-trained models whose output is combined with an unweighted average to produce the ensemble prediction of pixel-wise smoke labels.

3 RESULTS

We assessed the test set performance of models with different architectures and the two ensemble schemes and reported the results in Table 1. The best-performing architecture was PAN with the EfficientNet-b2 encoder, which was the architecture used in the initialization-based ensemble. Both the architecture-based ensemble and the initialization-based ensemble outperform the individual models across all IoU metrics.

Additionally, Figure 2 shows qualitatively that the ensemble predictions have smoother boundaries than the individual models, 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 compute uncertainties using the ensemble variability for every prediction.

Table 1: Test set IoU results across heavy, medium and light smoke density and over all densities with the single models and the two ensemble schemes. Encoder size refers to the size of the efficientnet encoder backbone used in the model. Gray shading indicates the models used in the architecture ensemble. Yellow shading indicates the ensemble schemes.

Model Architecture	Encoder Size	Heavy	Medium	Light	**All**
PAN [14]	2	0.349	0.478	0.664	**0.604**
DLV3P [13]	2	0.347	0.441	0.666	**0.599**
Unet++ [17]	1	0.369	0.472	0.654	**0.598**
Unet++ [17]	2	0.354	0.464	0.662	**0.597**
PSPNet [18]	2	0.374	0.482	0.651	**0.596**
DLV3P [13]	3	0.365	0.474	0.653	**0.595**
PAN [14]	3	0.324	0.461	0.658	**0.592**
PAN [14]	1	0.364	0.468	0.648	**0.590**
MANet [19]	2	0.352	0.478	0.646	**0.587**
LinkNet [20]	2	0.360	0.470	0.621	**0.570**
Architecture-based Ensemble	-	0.400	0.507	0.692	**0.635**
Initialization-based Ensemble	-	0.409	0.512	0.684	**0.631**

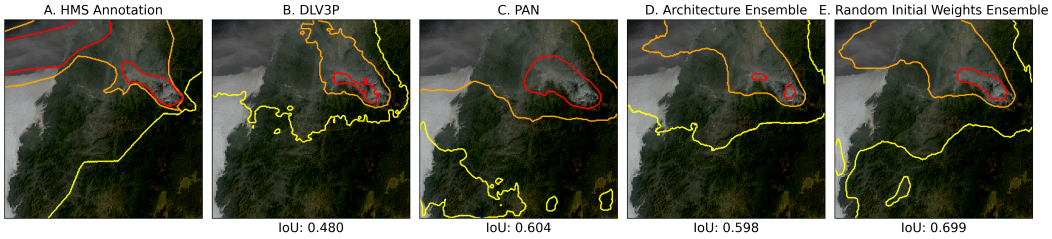

Figure 2: Example of smoke plume detection at (43.37°, -123.25°) on October 15 2022 15:50 UTC (within the test set). Red, orange, and yellow contours represent heavy, medium and light density smoke annotation, respectively. (A) is the ground truth annotation; (B-C) are two individual model predictions; (D) is the **architecture-based ensemble**; (E) is the **initialization-based ensemble**.

4 Limitations and Future Work

This proposal explores two schemes for building deep ensembles that both improve on test 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/type are. One area to explore is model stacking, where an optimized meta-model is used to combine multi-model outputs, as used in [9; 21; 22]. Furthermore, future work will quantify uncertainty in smoke annotations using the ensemble spread, enabling users like wildfire response teams and environmental agencies to assess the reliability of detections in real time. The code for this work is available at [23].

Acknowledgments

This research was supported by the NOAA Hollings Scholarship Program. All computation was performed on the NOAA Global Systems Laboratory Hera HPC system. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

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. doi: 10.1073/pnas.2011048118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2011048118.

[2] Wayne E. Cascio. Wildland fire smoke and human health. *The Science of the Total Environment*, 624:586–595, 05 2018. doi: 10.1016/j.scitotenv.2017.12.086. URL https://doi.org/10.1016/j.scitotenv.2017.12.086. Epub 2017 Dec 27.

[3] Jeff Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery. *ArXiv*, abs/2109.01637, 2021. URL https://api.semanticscholar.org/CorpusID:237416777.

[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. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2017.12.027. URL https://www.sciencedirect.com/science/article/pii/S003442571730603X.

[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. ISSN 1559-064X. doi: 10.1038/s41370-020-0246-y. URL https://doi.org/10.1038/s41370-020-0246-y.

[6] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

[7] 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. doi: 10.1080/02664763.2018.1441383. URL https://doi.org/10.1080/02664763.2018.1441383. PMID: 31631918.

[8] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001. ISSN 0262-8856. doi: https://doi.org/10.1016/S0262-8856(01)00045-2. URL https://www.sciencedirect.com/science/article/pii/S0262885601000452.

[9] 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. doi: 10.23919/EUSIPCO58844.2023.10289811.

[10] 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.

[11] 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.

[12] Rey Koki, Michael McCabe, Dhruv Kedar, Josh Myers-Dean, Annabel Wade, Jebb Q. Stewart, Christina Kumler-Bonfanti, and Jed Brown. Smokeviz: A large-scale satellite dataset for wildfire smoke detection and segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=NheuvQEWDt.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. URL https://arxiv.org/abs/1802.02611.

[14] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *CoRR*, abs/1805.10180, 2018. URL http://arxiv.org/abs/1805.10180.

[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. URL http://arxiv.org/abs/1807.10165.

[16] 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. doi: 10.1109/ICCRTEE64519.2025.11052908.

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

[18] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network, 2017. URL https://arxiv.org/abs/1612.01105.

[19] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. ISSN 1558-0644. doi: 10.1109/tgrs.2021.3093977. URL http://dx.doi.org/10.1109/TGRS.2021.3093977.

[20] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, page 1–4. IEEE, December 2017. doi: 10.1109/vcip.2017.8305148. URL http://dx.doi.org/10.1109/VCIP.2017.8305148.

[21] Linh Nguyen Van and Giha Lee. Optimizing stacked ensemble machine learning models for accurate wildfire severity mapping. *Remote Sensing*, 17(5), 2025. ISSN 2072-4292. doi: 10.3390/rs17050854. URL https://www.mdpi.com/2072-4292/17/5/854.

[22] 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. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2018.04.040. URL https://www.sciencedirect.com/science/article/pii/S0048969718311975.

[23] Annabel Wade. Dl ensemble smoke detection github repository, 2026. URL https://github.com/annabelwade/DL-ensembles-smoke-detection.

A Appendix

A.1 Data and code availability

The code for this work is openly available at [23]. The dataset can be accessed at https://noaa-gsl-experimental-pds.s3.amazonaws.com/index.html#SmokeViz/.

A.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 contributing to the diversity of the ensemble. Future work will investigate the variability of the individual model outputs and how that

relates to the ensemble performance, as well as how to optimize the ensemble design and size for best performance.

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.

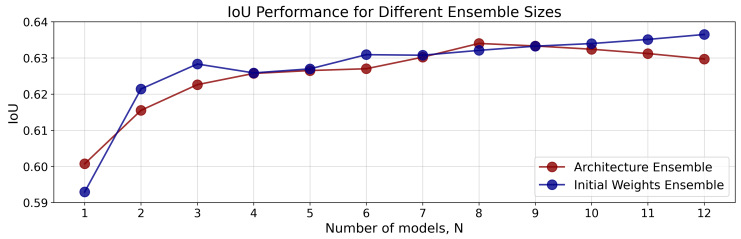

Figure 3: Overall IoU as a function of N for two ensemble design schemes: random initial weights (blue) and architecure-based (red).

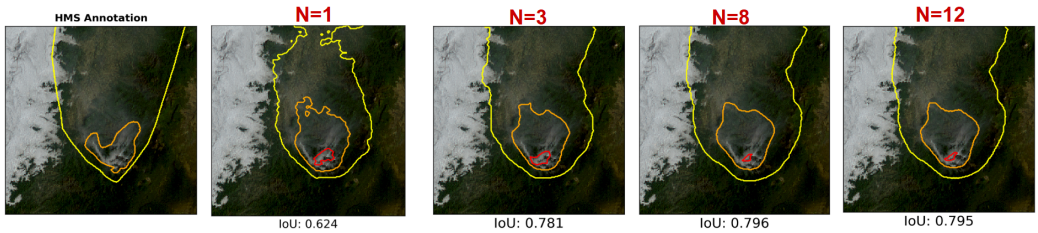

Figure 4: Example of smoke plume detection at (44.24, -122.74) on 2022/09/27 15:30 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. The first panel displays the ground truth HMS annotation;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

A.3 Intersection Over Union (IoU)

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

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

$$\text{IoU}_{\text{heavy}} = |y_{\text{heavy}} \cap y_{\text{heavy}}^*| \div |y_{\text{heavy}} \cup y_{\text{heavy}}^*| \tag{2}$$

$$\text{IoU}_{\text{medium}} = |y_{\text{medium}} \cap y_{\text{medium}}^*| \div |y_{\text{medium}} \cup y_{\text{medium}}^*| \tag{3}$$

$$\text{IoU}_{\text{light}} = |y_{\text{light}} \cap y_{\text{light}}^*| \div |y_{\text{light}} \cup y_{\text{light}}^*| \tag{4}$$