SAND2021-10685C

# Hyperspectral Image Target Detection Using Deep Ensembles for Robust Uncertainty Quantification

Rajeev Sahay[1,2], Daniel Ries[1], Joshua D. Zollweg[1], Christopher G. Brinton[2]
[1]Sandia National Laboratories, New Mexico, USA
[2]School of Electrical and Computer Engineering, Purdue University, USA
{rsahay,dries,jdzollw}@sandia.gov, cgb@purdue.edu

***Abstract*—Deep learning (DL) has been widely proposed for target detection in hyperspectral image (HSI) data. Yet, standard DL models produce point estimates at inference time, with no associated measure of uncertainty, which is vital in high-consequence HSI applications. In this work, we develop an uncertainty quantification (UQ) framework using deep ensemble (DE) learning, which builds upon the successes of DL-based HSI target detection, while simultaneously providing UQ metrics. Specifically, we train an ensemble of convolutional deep learning detection models using one spectral prototype at a particular time of day and atmospheric condition. We find that our proposed framework is capable of accurate target detection in additional atmospheric conditions and times of day despite not being exposed to them during training. Furthermore, in comparison to Bayesian Neural Networks, another DL based UQ approach, we find that DEs provide increased target detection performance while achieving comparable probabilities of detection at constant false alarm rates.**

***Index Terms*—deep learning, hyperspectral image processing, target detection, uncertainty quantification**

## I. Introduction

Target detection in hyperspectral images (HSI) consists of identifying the presence of one (or more) particular material(s) in an HSI's spectral bands. This type of detection is generally performed on a pixel-by-pixel basis, where all the spectral bands of a particular pixel are jointly analyzed for target detection. Prior work has employed likelihood based approaches [1], [2] for HSI target detection such as spectral matched filters, adaptive subspace detectors, and orthogonal subspace projection [3]. However, in addition to being computationally intensive at inference time, such methods typically assume the background to have a multivariate Gaussian distribution, which is often an inadequate representation of the underlying distribution, leading to lower target detection performance [2]. Additionally, if background scenes change, the covariance matrix obtained during training may not extrapolate well to new unseen backgrounds, reducing the generalizeability of such methods.

Recently, deep learning has been shown to achieve cutting edge target detection performance in HSI data without requiring an assumption on the background distribution of the image [4], [5]. Furthermore, deep learning detectors generalize better than likelihood based approaches in that deep learning detectors can perform accurate target detection in multiple atmospheric environments without being exposed to them during training. Yet, despite its impressive performance, standard deep learning models produce point estimates at inference time, with no measurement of uncertainty associated with the model's prediction. Uncertainty quantification (UQ) is, however, pivotal in high-consequence HSI processing applications, where the confidence of autonomous models, in detecting trace materials, needs to be trustworthy. As a result, deploying autonomous systems in high-consequence applications brings into question the trustworthiness of the model's output. Such trustworthiness is especially important in target detection when the abundance of a targeted material is very low, and the confidence associated with its detection needs to be very high.

In this work, we develop a robust HSI target detection framework, which incorporates UQ using deep ensembles (DE). We compare the efficacy of our framework to Markov Chain Monte Carlo (MCMC)-trained Bayesian Neural Networks (BNNs), which have also been used to incorporate UQ into HSI-based deep learning. Through our analysis, we find that both DEs and BNNs provide desirable characteristics in different inference scenarios.

Our problem setup is as follows: given an HSI scene, we would like to estimate the probability of each pixel containing a specific target material along with the associated uncertainty of that probability. We take a data-driven approach to this problem with the caveat that, similar to a real-world environment, we are required to learn from a limited number (as few as one) of spectral prototypes of a target material, which may vary between 0 - 100% in abundance per pixel. This constraint prevents our ensemble detectors from learning on different atmospheric environments and times of day; however, at test time, we evaluate the efficacy of our model on multiple atmospheres and times of day. The training and evaluation of our framework is further discussed in Sec. III.

## II. Methodology

Our proposed UQ framework for HSI target detection data builds upon prior successes in deep learning [6], [7]. Specifically, we train an ensemble of $B$ neural network models with identical architectures, $f : \mathbb{R}^k \rightarrow \mathbb{R}^2$, using the $N$ pixels from the available HSI prototype as inputs and binary labels (encoded in a one-hot representation) as targeted outputs, where target labels of 0 and 1 correspond to zero abundance and non-zero abundance of the target material, respectively, in the corresponding pixel.

HSI training data for target detection often has a large class imbalance, where there are overwhelmingly more non-target samples than target samples. As a result, detection algorithms often bias towards predicting pixels with trace amounts of target material as non-target samples at inference time. Recent studies have proposed computationally costly training algorithms for stronger detection performance on low-abundance pixels (e.g., [8], [9]) at inference time. Although such methods have been shown to be highly effective, we find that a relatively simple pixel extraction technique is sufficient to train a robust detector. Specifically, in an effort to restore class balance and improve target detection on low-abundance samples, while retaining enough samples for effective training, we extract all pixels from our training scene containing a non-zero target abundance and we randomly select twice the number of non-target samples to comprise the training set. As a result, $1/3$ of the training samples contain the target material and the remaining $2/3$ of the samples do not contain any amount of target material. This training method, as shown in Sec. IV, is effective for training both considered target detection model.

During training, we calculate the top $k$ functional principal components (fPCs) of the training data and project all testing data onto it for our evaluations. We find that, not only does using fPCs provide faster training times, but it also results in stronger overall performance compared to using the values from all the spectral bands of each pixel as input. Furthermore, we incorporate adversarial training into each detector to improve UQ performance by generating random perturbations from a Gaussian distribution on each fPC and augmenting these samples with the original unperturbed training data. Specifically, for each pixel's fPC representation, $x_i$, we generate $q$ corresponding noise vectors, $n_i \in \mathbb{R}^k$, where $n_i \sim \mathcal{N}(0, \sigma^2)$ (i.e., the distribution from which the Gaussian noise was generated had a variance of $\sigma^2$) such that the training set is augmented with $x_i + n_i$ for $i = 1, 2, 3, \ldots, q$. In this fashion, the final training set is comprised of $Nq$ samples.

At inference time, we use the distribution of predicted detection rates over the entire ensemble to determine the probability of detection as well as the prediction's associated uncertainty. We begin by calculating each input pixel's average probability of containing the target material by propagating the input through each detector in the deep ensemble. The mean probability of the input pixel containing the target material over the ensemble is given by

$$\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_i^b, \tag{1}$$

where $\hat{y}_i^b$ is the estimated probability of target, assigned by the $b^{\text{th}}$ model, of the $i^{\text{th}}$ pixel. From this, we calculate the confidence interval (CI) using the Normal approximation as justified by the Central Limit Theorem. The upper and lower bounds are given by

$$\hat{y}_i \pm z_\alpha \frac{s_i}{\sqrt{B}}, \tag{2}$$

where $s_i = \sqrt{\frac{\sum_{b=1}^{B} (y_i^b - \hat{y}_i)^2}{B-1}}$ is the standard deviation of the estimated probability that the $i^{\text{th}}$ pixel contains a target material, and $z_\alpha$ is the $1 - \alpha/2$ quantile of a Gaussian distribution. Eqn. (2) then gives the $(1-\alpha)\%$ CI. Here, for the purpose of UQ, we define a high confidence (HC) set, which contains pixels for which the model is 80% confident ($\alpha = 0.2$) that the pixel has either (i) at least an 80% chance of containing target, or (ii) at most a 20% chance of containing target. Note that the size of the HC set produced as a result of a particular model reflects that model's confidence in its predictions rather than its accuracy. For example, a model that results in a small proportion of samples in the HC set does not indicate the model is inaccurate and, similarly, a large HC set does not necessarily mean the model is correct in its predictions. We further analyze the proportion of HC sets in Sec. IV. In addition to the CI calculation, we also determine the area under the receiver operating characteristic (ROC) curve (AUC score), which quantifies the baseline performance of the DE, and the probability of detection (PD) at a constant false alarm rate (CFAR), which provides a specific UQ metric in addition to the CI.

## III. Implementation

**Dataset:** Our analysis is conducted on a synthetic dataset based on DIRSIG Megascene, from which we simulated nine different HSI scenes using combinations of three different MODTRAN-based atmospheres (Mid-Latitude Summer (MLS), Sub-Arctic Summer (SAS), and Tropical (TROP)) and three different times of day (1200, 1430, and 1545). Further details regarding the dataset generation and processing can be found in [10]. To serve as targets, we manually inserted green discs randomly through each scene such that some targets filled multiple pixels while others filled a small fraction of a pixel. In this fashion, the abundance level the target material (green paint) varied between 0 – 100% per pixel. Note that we are interested in detecting and quantifying the associated uncertainty of the presence of target in a pixel, regardless of abundance level. However, we present our results by abundance level to evaluate the performance of our method at varying target abundance levels. For example, the probability of detection (at a particular CFAR) is more interesting to consider on low target abundance levels, where higher confidence associated with the DE's prediction is desired. Fig. 1 shows a pseudo color rendering for the MLS atmosphere at 1200, which was the scene we used for training.

**Training Details:** For training, the top $k = 25$ fPCs of a subset of pixels from the left half of MLS 1200 were used for training the DE, and the fPCs of all pixels from the right half of all nine scenes were used as test sets. Note that the top $k = 25$ fPCs accounted for 99.999% of the variability of spectral reflectance in the training dataset. We implemented adversarial training using $q = 30$ and $\sigma = 0.001$. Note that our adversarial training method relies on random Gaussian perturbations as opposed to gradient-based adversarial training (as in [6]), which relies on the gradient of the trained model to generate adversarial samples for re-training. Furthermore, by

**Fig. 1:** Pseudo color render of Megascene MLS 1200.

**TABLE I:** Proportion of pixels captured in HC set on each scene for each model.

| Scene | BNN | DNN DE | CNN DE |
|---|---|---|---|
| MLS 1200 | 0.81 | 0.99 | 0.91 |
| MLS 1430 | 0.79 | 0.99 | 0.90 |
| MLS 1500 | 0.74 | 0.98 | 0.90 |
| SAS 1200 | 0.66 | 0.99 | 0.90 |
| SAS 1430 | 0.67 | 0.99 | 0.90 |
| SAS 1545 | 0.88 | 0.99 | 0.89 |
| TROP 1200 | 0.51 | 0.97 | 0.91 |
| TROP 1430 | 0.50 | 0.96 | 0.91 |
| TROP 1545 | 0.53 | 0.96 | 0.90 |

only training on one scene at a particular time and atmosphere, we are able to understand the model's ability to detect targets in scenes it has never seen before. This is particularly important for our application, since we cannot realistically expect to have training data in all atmospheres and times of day in practice.

**Model Architectures:** We considered two different deep learning architectures to constitute the DE: (i) a fully connected dense neural network (DNN DE) consisting of three layers with 10 sigmoid units each (note that this architecture was also used for the BNN) and (ii) a convolutional neural network (CNN DE) with two convolutional ReLU layers, consisting of 64 feature maps and a $4 \times 1$ kernel and 32 feature maps and a $3 \times 1$ kernel, respectively, followed by a 128-unit dense ReLU layer. Each DE used $B = 10$ models. The BNN had $\mathcal{N}(0, 10)$ priors on all weights and used MCMC to train. Standard metrics (e.g., training and validation loss) indicated model convergence of predictions during training.

## IV. Results and Evaluation

We begin by evaluating the detection performance of our considered models. Fig 2 shows the ROC curves and AUC scores for target detection using both DEs and BNNs on the SAS 1430 scene at different target abundance levels. Note that this scene contains a different atmosphere and is captured at a different time of day compared to the training scene, but the target material (green paint) is the same. Here, we see that DEs outperform BNNs for target detection overall, and especially for low target abundance levels. In particular, when the target abundance is less than or equal to 1%, the DNN DE and CNN DE obtain improvements in AUC of 0.09 and 0.11, respectively, over the BNN. At higher abundance levels, the performance of each model is roughly equivalent, which is expected since higher abundance targets are more easily detected.

We now evaluate the UQ performance of each considered model. In Table I, we show the proportion of samples captured in the HC set of each model. We see that the BNN attains a relatively smaller HC set on scenes with different atmospheres and times of day compared to the training scene. In contrast, we see that the DNN DE captures almost every sample in nearly every scene in its HC set. Although a high portion of samples are captured in the HC set of the DNN DE, we believe that this may indicate that it is overconfident in its predictions. We find a model that produces a narrow interval about its predicted distribution is often not much different from a standard deep learning detector that produces point estimates at its output. In this regard, a model producing a large HC set, such as the DNN DE, fails to provide strong UQ characteristics. On the other hand, the CNN DE provides a reasonable HC set while also attaining robust detection performance.

The aforementioned trends and insights are further validated in Figs. 3 - 6, where we show the PD at a CFAR of 1% and 0.1% averaged across all nine scenes, in which the DNN DE does not outperform the BNN, in terms of PD, across all abundance levels despite having a larger HC set. Furthermore, we find that the probability of detection across all target abundance levels are nearly equivalent between the CNN DE and the BNN, as shown in Figs. 4 and 6, despite the CNN DE capturing significantly more samples in its HC set. As shown in Figs. 3 and 5, the PD at low CFAR levels does not outperform the baseline BNN in either the full testing sets or the HC sets. This reinforces our finding that, despite having a large HC set (as shown in Table I), the DNN DE does provide strong UQ metrics. Thus, we find the CNN DE to be the most desirable detection framework due to its baseline performance and UQ performance.

## V. Conclusion

Deep learning models have achieved cutting edge target detection performance in hyperspectral images. Yet, standard deep learning models produce point estimates at test time, with no measure of uncertainty associated with the model's prediction. In this work, we presented a deep ensemble detection framework, which leveraged the robustness of standard deep learning models, while simultaneously providing uncertainty quantification metrics. In this capacity, we found that, in comparison to Bayesian neural networks, deep ensembles attain

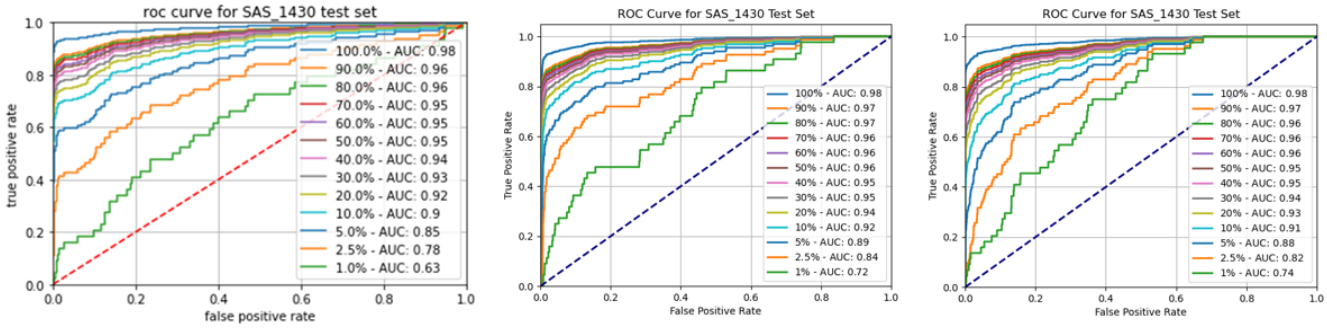


**Fig. 2:** ROC curves for different target abundance levels on BNN (left), DNN DE (middle), and CNN DE (right). Each DE achieves a higher AUC than the BNN on lower abundance levels, while the performance of all three models is roughly equivalent at higher abundance levels.

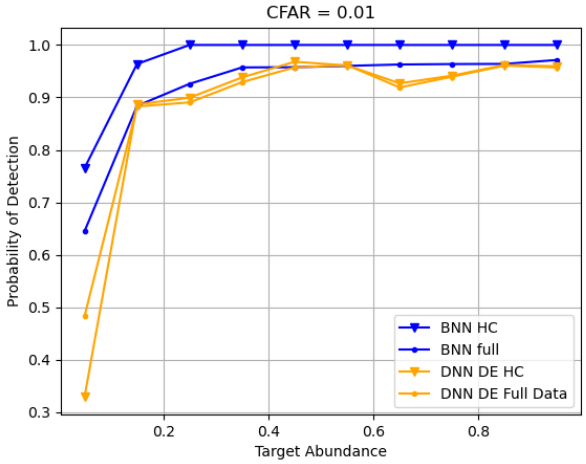


**Fig. 3:** Comparison between BNN and DNN DE of PD at CFAR = 0.01 (averaged across all nine scenes). The BNN delivers a higher PD on the HC set, while the PD on the full set is somewhat more equivalent between the BNN and the DNN DE.

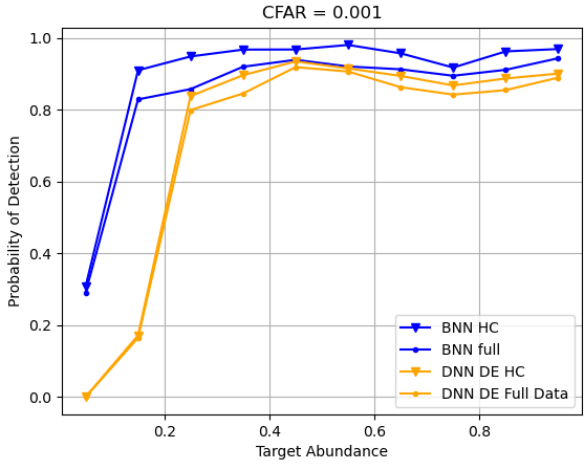


**Fig. 5:** Comparison between BNN and DNN DE of PD at CFAR = 0.001 (averaged across all nine scenes). The BNN achieves a higher PD on both the HC and full set at all target abundance levels, with a greater PD at lower abundance levels.

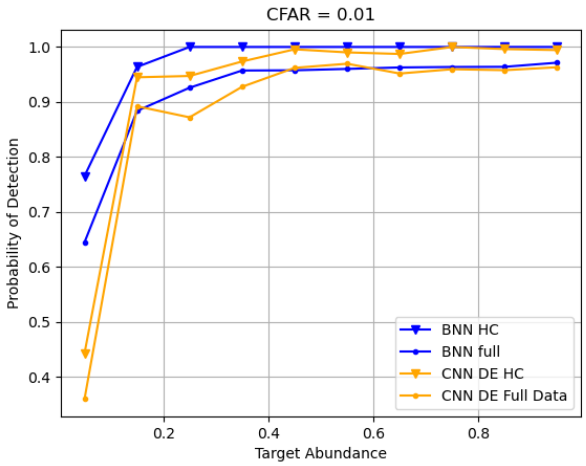


**Fig. 4:** comparison between BNN and CNN DE of PD at CFAR = 0.01 (averaged across all nine scenes). At lower abundance levels, the BNN obtains a higher averaged PD than the CNN DE, but the PD becomes roughly equivalent at higher abundance levels.

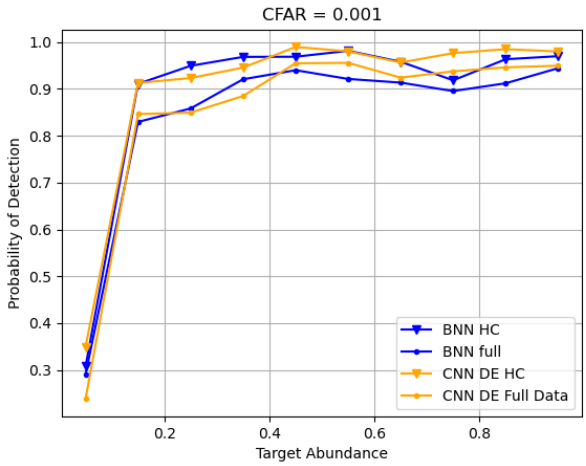


**Fig. 6:** Comparison between BNN and CNN DE of PD at CFAR = 0.001 (averaged across all nine scenes). The higher PD, on both the HC and full set, between the BNN and CNN DE depends on the abundance level of the target material.

higher detection performance while delivering higher degrees of confidence with its predictions. Furthermore, we found that the deep learning architecture used for constructing the DE detectors is pivotal in attaining useful uncertainty quantification metrics. For example, a deep ensemble constructed of an entirely dense architecture provides overly confident predictions, making their uncertainty quantification performance equivalent to employing a single detection model. On the other hand, convolutional architectures provide a better balance between detection performance and uncertainty quantification. Finally, we found that at low target abundance levels, the Bayesian neural network achieves only slightly higher probabilities of detection on its high confidence samples despite having significantly smaller high-confidence sets in comparison to the CNN deep ensemble. Thus, we find that the convolutional deep ensemble provides the most desirable characteristics in terms of both detection performance and uncertainty quantification metrics.

## ACKNOWLEDGMENT

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

## REFERENCES

[1] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[2] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, 2002.

[3] N. M. Nasrabadi, "Hyperspectral target detection : An overview of current and future challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 34–44, 2014.

[4] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5408–5423, 2018.

[5] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in NeurIPS*, 2017.

[7] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton, "Hyperparameter ensembles for robustness and uncertainty quantification," in *Advances in NeurIPS*, 2020.

[8] G. Zhang, S. Zhao, W. Li, Q. Du, Q. Ran, and R. Tao, "Htd-net: A deep convolutional neural network for target detection in hyperspectral imagery," *Remote Sensing*, vol. 12, no. 9, 2020.

[9] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 844–853, 2017.

[10] D. Anderson, J. Zollweg, and B. Smith, "Paired neural networks for hyperspectral target detection," in *SPIE Application of Machine Learning*, 2020.