Towards Scalable Identification of Brick Kilns from Satellite Imagery with Active Learning

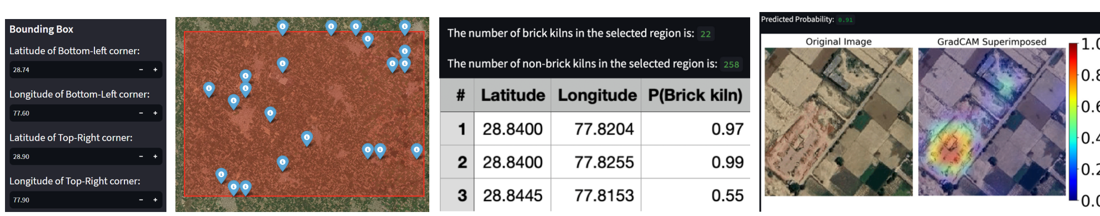
Aditi Agarwal, Suraj Jaiswal, Madhav Kanda, Dhruv Patel, Rishabh Mondal, Vannsh Jani, Zeel B Patel, Nipun Batra
Indian Institute of Technology, Gandhinagar

Sarath Guttikunda
Urban Emissions

Abstract

Air pollution is a leading cause of death globally, especially in south-east Asia. Brick production contributes significantly to air pollution. However, unlike other sources such as power plants, brick production is unregulated and thus hard to monitor. Traditional survey-based methods for kiln identification are time and resource-intensive. Similarly, it is time-consuming for air quality experts to annotate satellite imagery manually. Recently, computer vision machine learning models have helped reduce labeling costs, but they need sufficiently large labeled imagery. In this paper, we propose scalable methods using *active learning* to accurately detect brick kilns with minimal manual labeling effort. Through this work, we have identified more than 700 new brick kilns across the Indo-Gangetic region: a highly populous and polluted region spanning 0.4 million square kilometers in India. In addition, we have deployed our model as a web application for automatically identifying brick kilns given a specific area by the user.

Keywords: Active Learning, Satellite Imagery, Sustainable Development, Air Pollution



(a) (b) (c) (d)

Figure 1: Screenshots from our web application that help detect brick kilns (a) Selecting the coordinates of the bounding box (red rectangle) (b) Markers in the bounding box where the model predicts the existence of a brick kiln (c) Statistics of number of brick kilns detected, their coordinates and model's predicted probabilities (d) Grad-CAM (Selvaraju et al., 2019) visual showing where our model focuses on predicted brick kiln image (Best viewed in color)

1. Introduction

Air pollution kills seven million people worldwide, and 22% of casualties are only from India (UNEP, 2019). Annual average $PM_{2.5}$ (Particulate matter of size ≤ 2.5 μm) of India was 24 $\mu g/m^3$ in 2020, which is significantly higher than the annual WHO limit of 5 $\mu g/m^3$ (Guttikunda and Nishadh, 2022). Air quality researchers use physics-based simulators such as CAMx[1] to model the air quality (Guttikunda et al., 2019) using an inventory of major sources.

1. `https://www.camx.com/`

Brick kilns are one such major source of pollution. They contribute up to 91% of air pollution in South Asia (WorldBank, 2020). Also, in South Asia, 144,000 units of brick kilns produce 0.94M tonnes of PM, 3.9M tonnes of CO, and 127M tonnes of CO_2 annually employing 15M workers (Rajarathnam et al., 2014).

Monitoring these small, unregulated kilns using traditional survey methods is labor and resource-intensive and lacks scalability for maintaining a dynamic inventory. Air quality experts who run physical models such as CAMx leverage satellite imagery to detect these kilns using manual annotation. Scaling this for a country like India would require manual labeling of millions of images, requiring years of time due to the sparsity of brick kilns. Recent studies (Lee et al., 2021) have leveraged popular pretained CNN models like VGG16 (Huang et al., 2018) and ResNet (He et al., 2015) for transfer learning-based identification of brick kilns with imagery from a private satellite. However, such methods require extensive human annotation and expertise to curate a vast dataset.

Our paper proposes scalable methods for identifying brick kilns using publically available satellite imagery. We propose to leverage active learning (Settles, 2009) to strategically curate a dataset for any new region. We also leverage pretrained CNN models like VGG16 (Huang et al., 2018) and ResNet (He et al., 2015) and fine tune them on our dataset. We utilize Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) to obtain uncertainty. We show that using our methods, we need to annotate only a small number of images to obtain brick kiln locations in a new region. On performing active learning on the Indian dataset, we concluded that we needed 70% fewer samples than random to achieve a similar F1 score. We also find that we could reach 97% of optimal F1 score with active learning, whereas random could reach only 90% with the same number of samples labeled.

Finally, we have developed a web application [2] offering users an accessible and interactive interface for brick kiln detection in a given region of interest. Figure 1 shows our web application which takes in bounding boxes of the area of interest and detects the kilns present in the region while also showing Grad-CAM (Selvaraju et al., 2019) visuals to highlight the focus area of the model. Our work is fully reproducible, and we intend to release the scripts and data upon acceptance.

2. Dataset

We use two different datasets for this work: dataset released by (Lee et al., 2021) from Bangladesh as shown in Figure 2a; and our own curated dataset from Delhi, India shown in Figure 2b. With the help of researchers from (Guttikunda et al., 2019), we curated a first of its kind dataset consisting of 762 brick kiln images across Delhi, India shown in Figure 2c. These images are of size 256×256, taken using Google Static Maps API at zoom level 17 with a 1-meter-to-pixel ratio to match the configuration of the Bangladesh (Lee et al., 2021) dataset. We have specifically curated the images Delhi, India since it is a highly populous region characterised with alarming levels of air pollution. Additionally, this region is located in the highly fertile Indo-Gangetic plain which it a hotspot for production of bricks. The dataset also contains 2000 non-brick kiln images from structures visually similar to brick kilns to make the dataset more challenging and our model robust. These

2. https://brick-kilns-detector.streamlit.app/

include farms, barren land, and thermal power plants taken from the same region. A team of three annotators manually verified all images independently to exclude brick kiln images. The Cohen-Kappa score (McHugh, 2012) of the annotators is 99.33%.

(a) (b) (c)

Figure 2: Satellite images of brick kilns from (a) Bangladesh and (b) India. Image (c) presents brick kilns in the Delhi, India dataset curated with the help of an air quality expert. All the kilns lie on the outskirts of Delhi.

3. Approach

Air quality experts perform manual annotation on satellite imagery to identify brick kilns. However, due to the sparsity of brick kilns it is challenging to scale these methods for a large area. Recently, computer vision machine learning models have helped reduce the labeling costs. However, they still need sufficiently large labelled data. Our approach aims to leverage active learning to strategically curate a dataset for any new region with substantially lower number of annotations.

3.1 Modeling

We use a variety of pre-trained Convolutional Neural Network (CNN) models, which include VGG16 (Huang et al., 2018), ResNet50 (He et al., 2015), DenseNet121 (Huang et al., 2018) and EfficientNet-B0 (Tan and Le, 2020), with pre-trained ImageNet weights.

1. **Zero-Shot learning**: We finetune the pretrained models on the Bangladesh dataset (Lee et al., 2021) and evaluate its performace on the Indian dataset.
2. **Fine Tuned on Target Region**: We finetune the pretrained models on Indian dataset and evaluate the model performance on Indian dataset.

3.2 Obtaining model uncertainty

We use MC Dropout (Gal and Ghahramani, 2016), the state-of-the-art and computationally effective method, to obtain predictive uncertainties. Essentially, it does multiple forward passes (get MC samples) through the model while keeping the dropout layer active with different random seeds. We then obtain mean and standard deviation across these MC samples to get a predictive distribution. Refer to appendix A.1.1 for more details.

3.3 Active Learning

Active learning is a strategy to intelligently query samples that improve the model the most. We use an acquisition function to choose which samples to label for improving the model. We now discuss the baselines and acquisition strategies used in our experiments from (Gal et al., 2017) [3]. We also propose a new acquisition strategy to combat the class imbalance problem.

1. **Entropy**: Entropy (Shannon, 1948) is a measure of model's uncertainty. It might be useful to label the points where the predictive entropy is the highest. Entropy is defined as:
$$\mathbb{H}[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] = -\sum_{c} p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}}) \log p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}}) \tag{1}$$
where $p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ is predicted probability for class c.

2. **BALD**: Bayesian Active Learning by Disagreement (Gal et al., 2017) is a method to maximise the information gained about the model parameters, i.e. maximise the mutual information between predictions and model posterior.It is mathematically defined as:
$$\mathbb{I}[y, \boldsymbol{\theta} \mid \mathbf{x}, \mathcal{D}_{\text{train}}] = \mathbb{H}[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] - \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})}[\mathbb{H}[y \mid \mathbf{x}, \boldsymbol{\theta}]]$$
with $\boldsymbol{\theta}$ the model parameters and $\mathbb{H}[y \mid \mathbf{x}, \boldsymbol{\theta}]$ is the entropy of y given model weights $\boldsymbol{\theta}$.

3. **Subset Scoring**: We propose a new acquisition function to explicitly select the subset of images classified as brick kilns in each iteration. The intuition is to select the points that are predicted as positive, but the model is not confident about them. Including such points may boost model's performance for the positive class especially in case of class imbalance. The function is defined as follows:
$$\mathbb{S}[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] = \mathbb{I}(\hat{y} = c) \cdot \alpha[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] \tag{2}$$
where α can be one of the acquisition functions discussed earlier.

4. **Random**: This acquisition function is equivalent to choosing an image uniformly at random from the pool dataset. Prior literatures (Gal et al., 2017; Settles, 2009) on active learning consider random sampling as the baseline.

5. **Total baseline**: We consider this an oracle baseline, where we train the model on the entire labeled dataset except a hold-out test dataset and evaluate the performance on the test dataset.

4. Evaluation

We first describe the three main experiments:

1. First, we evaluate the performance of zero shot learning where we fine tune the pre-trained model on the Bangladesh dataset and test on the Indian dataset.

3. (Gal et al., 2017) suggests numerous acquisition strategies like maximising the variation ratios and maximising the mean of standard deviation. However, for a binary classification task all the strategies behave similar to BALD

2. Second, we evaluate the performance of pretrained models fine-tuned on the Indian dataset. This is the same as the total baseline mentioned in Section 3.3 for the Active Learning experiment.
3. Third, we perform active learning using different acquisition functions and evaluate the need for labelled data.

4.1 Experimental setup

We first discuss the experimental settings common across the three experiments. We divide the Indian dataset into a (80%, 20%) stratified split such that the train set and test set have an equal proportion of brick kilns i.e., 610 and 152 brick kiln images in train and test set, respectively. This 20% split is used as a test set across all our experiments. For the first experiment (zero-shot performance), we fine-tune (till convergence) on the entire Bangladesh dataset containing 2804 images and test on the 576 Indian test images. For the second experiment, we fine-tune (till convergence) on the 80% train Indian dataset(2209 images). For the third experiment, we further split this 80% train set into a stratified 1:99 ratio set which we use for training (22 images) and as pool set (2187 images) respectively for this experiment. We initially fine-tune the pre-trained model on the 22 images. Then, in the active learning loop, we add a single image per active learning iteration and fine-tune for 5 epochs. To compare the performances of our models, we use standard metrics used in prior literature: accuracy, precision, recall, and F1 score.

4.2 Results and Discussion

Models	Zero Shot Learning			Fine tuned on Indian dataset (Total Baseline)		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
VGG16	0.92	0.45	0.60	1.00	0.94	0.97
ResNet50	0.96	0.68	0.80	**1.00**	**0.95**	**0.97**
DenseNet121	0.93	**0.71**	**0.81**	0.99	0.95	0.97
EfficientNetB0	**1.00**	0.66	0.79	0.98	0.96	0.97

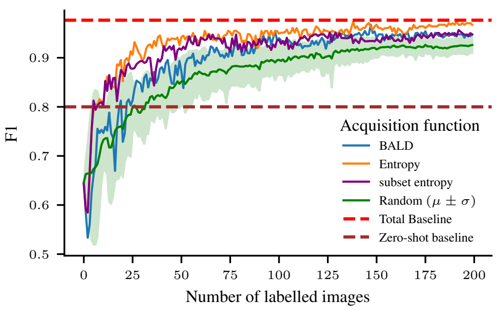
Table 2: Performance metrics for different models on the Indian dataset in *Zero-shot* setting where models are trained only on the Augmented Bangladesh dataset v/s models fine-tuned on the Indian dataset. It is evident from the results that models fine-tuned on the Indian dataset have higher metric values.

4.2.1 Zero Shot Learning v/s Fine tuned on Indian dataset

We show the result for zero-shot learning in Table 2. Different models are able to achieve an F1 score close to 0.8. This shows that we can extend the Bangladesh model to any country for which the model is not explicitly trained and reduce the manual efforts. On the contrary, most models have low recall due to a lack of adaptability for a new region. The models that are fine-tuned on the Indian dataset in contrast achieve higher metrics overall. Based on the performance of the metric scores obtained above, we select ResNet for performing active learning.

4.2.2 Active Learning

To compare the performance achieved in active learning across different acquisition strategies, we compare the metric scores after each iteration. We had seen previously from Table 2 ResNet50 achieves 0.97 F1 score, which is an upper bound for the active learning experiments, as these models have been trained on the entire train + pool dataset. Figure 3a shows that active learning-based acquisition functions perform favourably with respect to the random baseline. Our proposed **Subset Entropy** baseline gives the best performance when only a small number of images have been labelled. Table 3b shows that we can get within 5% of the best achievable performance using a significantly lower number of images for different acquisition functions when compared against the random baseline.



(a)

%	BALD	Entropy	Subset Entropy	Subset BALD	Random
25%	7	4	4	6	17
20%	13	5	5	11	32
15%	26	11	11	24	49
10%	35	17	13	32	78
5%	65	33	33	61	200

(b)

Figure 3: (a) F1 score v/s number of labeled images for different querying strategies. Our proposed acquisition 'subset entropy' performs the best in the initial iterations of active learning and is always better than state-of-the-art BALD acquisition. (b) The table presents the number of images needed to label to attain F1 scores within 5%, 10%, 15%, 20%, 25% of the total baseline.

5. Limitations and Future Work

- In our current work, we only looked at the binary classification task. Drawing inspiration from (Lee et al., 2021), we plan to additionally localize the kilns in the image and extend our active learning pipeline towards multiple objectives: localization and classification.
- Our current work treated the classification problem as a binary classification task. In the future, we plan to study this formulation as a one-class task. Correspondingly, we also plan to look at specialized losses such as the focal losses (Lin et al., 2017).

6. Conclusion

Our goal was to develop a scalable method to detect brick kilns. We conclude from our results that satellite data can be used to detect brick kilns accurately. Further, we conclude that we can develop accurate models by actively annotating images from the target region. We believe that our work will likely benefit key stakeholders such as scientists building emission inventories and policy makers looking at regulating and monitoring brick kilns by automating the current manual process of mapping brick kilns.

References

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50:5–43, 2003.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, 2017.

Sarath Guttikunda and KA Nishadh. Evolution of india's pm 2.5 pollution between 1998 and 2020 using global reanalysis fields coupled with satellite observations and fuel consumption patterns. *Environmental Science: Atmospheres*, 2(6):1502–1515, 2022.

Sarath K Guttikunda, KA Nishadh, Sudhir Gota, Pratima Singh, Arijit Chanda, Puja Jawahar, and Jai Asundi. Air quality, emissions, and source contributions analysis for the greater bengaluru region of india. *Atmospheric Pollution Research*, 10(3):941–953, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

Jihyeon Lee, Nina R Brooks, Fahim Tajwar, Marshall Burke, Stefano Ermon, David B Lobell, Debashish Biswas, and Stephen P Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118 (17):e2018863118, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

M. L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

Uma Rajarathnam, Vasudev Athalye, Santhosh Ragavan, Sameer Maithel, Dheeraj Lalchandani, Sonal Kumar, Ellen Baum, Cheryl Weyant, and Tami Bond. Assessment of air pollutant emissions from brick kilns. *Atmospheric Environment*, 98:549–553, 2014. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2014.08.075. URL https://www.sciencedirect.com/science/article/pii/S1352231014006888.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7. URL `https://doi.org/10.1007%2Fs11263-019-01228-7`.

Burr Settles. Active learning literature survey. 2009.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

UNEP. Emissions gap report 2019, 11 2019.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

WorldBank. Dirty stacks, high stakes: An overview of brick sector in south asia, 2020.

Appendix A.

A.1 Background

We now discuss background work across: obtaining uncertainty from neural networks, active learning, and different acquisition strategies.

A.1.1 Uncertainty in Neural Networks

Neural networks have been shown to be overconfident, i.e. they can assign a high probability to a wrong label (Wilson and Izmailov, 2020). Bayesian neural networks (BNNs) (Wilson and Izmailov, 2020; MacKay, 1992) can provide better uncertainty estimates by accounting for uncertainty in the model parameters. BNNs introduce a prior $p(\boldsymbol{\theta})$ on model parameters $\boldsymbol{\theta}$. After observing the data ($\mathcal{D}$), we can get the conditional distribution (posterior) over the parameters ($p(\boldsymbol{\theta}|\mathcal{D})$). The conventional way of predicting from BNNs is to use MCMC (Markov Chain Monte Carlo) (Andrieu et al., 2003) methods which are slow (Blundell et al., 2015). Monte Carlo dropout (MC dropout) (Gal and Ghahramani, 2016) has emerged as an efficient modern technique for estimating uncertainty within neural networks. In the MC dropout method, we run multiple forward passes over the input by randomly dropping the weights or applying dropout (Srivastava et al., 2014). The authors provide theoretical guarantees for MC dropout as an approximate Bayesian method.

A.1.2 Active Learning

The efficacy of deep learning models relies on the availability of labeled training data, which often demands extensive manual annotation efforts. This challenge has prompted the exploration of active learning (Settles, 2009) techniques as a strategic approach to minimize annotation costs while retaining model performance. Active learning is a strategy to intelligently query samples that improve the model the most. We use an acquisition function to choose the samples and pass them to a human annotator or any source that can label them. Following is the algorithm of active learning:

1. We train our model on the initial dataset $\mathcal{D}_{\text{train}}$.
2. We evaluate the acquisition function on the unlabeled data points $\mathcal{D}_{\text{pool}}$ (pool data) and label the points which optimize the acquisition function.
3. We add the newly labeled points into the initial dataset and retrain/fine-tune the model.
4. We continue the steps 2 and 3 for K iterations, or till we get sufficiently better performance on validation data.

Now we discuss the acquisition functions used in our work:

entropy: entropy (Shannon, 1948) is a measure of model's uncertainty. It might be useful to label the points where the model is most uncertain. entropy is defined as:

$$\mathbb{H}[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] = -\sum_{c} p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}}) \log p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}})$$

where $p(y = c \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ is predicted probability for class c.

BALD: BALD or Bayesian Active Learning by Disagreement (Gal et al., 2017) is a method to choose points where the model is overall uncertain while computing the entropy over mean predictions of MC dropout samples, but each MC dropout prediction is highly confident about the class prediction. It is mathematically defined as:

$$\mathbb{I}[y, \boldsymbol{\theta} \mid \mathbf{x}, \mathcal{D}_{\text{train}}] = \mathbb{H}[y \mid \mathbf{x}, \mathcal{D}_{\text{train}}] - \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})}[\mathbb{H}[y \mid \mathbf{x}, \boldsymbol{\theta}]]$$

with $\boldsymbol{\theta}$ the model parameters and $\mathbb{H}[y \mid \mathbf{x}, \boldsymbol{\theta}]$ is the entropy of y given model weights $\boldsymbol{\theta}$.

A.2 Model evaluation

Training and Testing on Bangladesh Dataset We create stratified splits of the dataset and have 1748 images as the training set, 438 images as the validation set, and 618 images as the test set. We use early stopping, based on validation loss, to avoid overfitting the model. We fix the learning rate for each model to 2×10^{-5}. Table A indicate the metrics for each model.

Model	Precision	Recall	F1 Score
VGG16	0.816	0.794	0.805
ResNet50	**0.880**	**0.808**	**0.842**
DenseNet121	0.864	0.780	0.820
EfficientNetB0	0.865	0.795	0.829

Table A: Performance metrics for different models using augmented Bangladesh dataset for training and testing. We use the same train-test split provided in the paper (Lee et al., 2021). We find that ResNet50, DenseNet121, and EfficientNetB0 models have comparable performance.

Table B and Table C are extended version of Table A and Table 2. These tables include results for the augmented and non-augmented versions of training data. We observe that the augmented version is most of the time better than the non-augmented version in terms of scoring metrics.

Model	Accuracy	Precision	Recall	F1 Score
VGG16	0.948	0.753	**0.835**	0.792
VGG16-A	0.955	0.816	0.794	0.805
ResNet50	**0.956**	**0.883**	0.726	**0.797**
ResNet50-A	**0.964**	**0.880**	**0.808**	**0.842**
DenseNet121	0.948	0.830	0.698	0.761
DenseNet121-A	0.959	0.864	0.780	0.820
EfficientNetB0	0.943	0.788	0.712	0.748
EfficientNetB0-A	0.961	0.865	0.795	0.829

Table B: Performance metrics for different vanilla and augmented models using Bangladesh dataset for training and testing.

Model	Accuracy	Precision	Recall	F1 Score
VGG16	**0.893**	0.805	**0.810**	**0.807**
VGG16-A	0.837	0.920	0.451	0.605
ResNet50	0.863	0.943	0.536	0.683
ResNet50-A	0.906	0.963	0.682	0.801
DenseNet121	0.840	**0.985**	0.431	0.600
DenseNet121-A	**0.907**	0.939	**0.712**	**0.810**
EfficientNetB0	0.870	0.966	0.549	0.700
EfficientNetB0-A	0.905	**1.000**	0.660	0.795

Table C: Performance metrics for different vanilla and augmented models using Bangladesh for training and Indian dataset for testing

A.3 Expanding the brick kiln dataset spatially

We map brick kilns across the Indo-Gangetic (IG) plain, which covers 14 Indian states along the river Ganges. A significant portion of the Indian population resides in the IG plain. The IG plain is also likely to house a large number of brick kilns due to the availability of favorable soil conditions.

The IG plain covers approximately 0.4 million sq. km, equivalent to 6.4 million images at the current resolution. In an ideal scenario, we would perform a forward pass of our model on all these images to obtain their labels. However, owing to the limited API calls to access these satellite images, we planned to use 82,000 images instead. randomly sampling the IG plain for selecting 82,000 images may be inefficient as it misses two important domain insights: i) there is spatial locality to brick kilns, i.e., we usually have a cluster of brick kilns; ii) the brick kiln sites are more likely to be present close to the river banks. Thus, to create this IG plain dataset of 82,000 images, we first manually identified 189 brick kiln sites. We then looked into neighboring images and expanded our dataset to include 82,000 images. We then ran the forward pass on these 82,000 images. Our model classified 1847 of them as brick kilns. We looked into all these images and identified new 704 images containing brick kilns. These are shown in Figure A. For the non-classified images, we randomly sampled 1000 images, and after manual inspection, we found 996 of them to be correctly classified. Thus through our approach we were able to reduce the annotation

A.4 Deployment

We deploy a web application on Streamlit, as depicted in Figure 1, offering users an accessible and interactive interface for brick kiln detection in a given area of interest. Once the bounding box is defined, our model identifies brick kilns within this area and provides the coordinates of the brick kilns. Grad-CAM (Selvaraju et al., 2019) visuals accompany these on the original brick kiln image to highlight the areas where the model focuses.

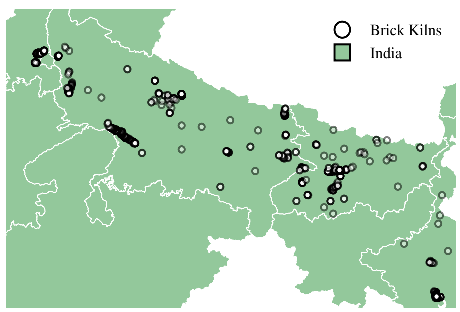


Figure A: We initially manually located approximately 189 brick kilns in the Indo-Gangetic plain. Subsequently, our model automatically detected an additional 704 new brick kilns in the vicinity of the manually identified ones, as illustrated in the figure