



OPEN

Brick Kiln Dataset for Pakistan's IGP Region Using AI

DATA DESCRIPTOR

Muhammad Suleman Ali Hamdani¹, Khizer Zakir², Neetu Kushwaha³, Syeda Eman Fatima¹ ¹ & Hassan Aftab Sheikh³

Brick kilns are a major source of air pollution in Pakistan, with many operating without regulation. A key challenge in Pakistan and across the Indo-Gangetic Plain is the limited air quality monitoring and lack of transparent data on pollution sources. To address this, we present a two-fold AI approach that combines low-resolution Sentinel-2 and high-resolution imagery to map brick kiln locations. Our process begins with a low-resolution analysis, followed by a post-processing step to reduce false positives, minimizing the need for extensive high-resolution imagery. This analysis initially identified 20,000 potential brick kilns, with high-resolution validation confirming around 11,000 kilns. The dataset also distinguishes between Fixed Chimney and Zigzag kilns, enabling more accurate pollution estimates for each type. Our approach demonstrates how combining satellite imagery with AI can effectively detect specific polluting sources. This dataset provides regulators with insights into brick kiln pollution, supporting interventions for unregistered kilns and actions during high pollution episodes.

Background & Summary

Air pollution in South Asia is responsible for over 2 million premature deaths annually with pollution levels exceeding World Health Organization (WHO) air quality standards by up to ten times for particulate matter (PM_{2.5})^{1,2}. The Indo-Gangetic Plain (IGP), a region (Fig. 1a) encompassing 13.5 million hectares of land across India, Pakistan, and Bangladesh is one of the most polluted regions globally^{3,4}. The region's meteorological and topographical region restricts the dispersion of pollutants, making it a global air pollution hotspot⁵. This is also highlighted from satellite data observations that have shown a temporal increase from 2011 to 2021 in pollutants such as sulphur dioxide (SO₂), nitrogen oxides (NO_x), and other aerosols⁶.

In developing countries, air pollution from informal sectors is often unregulated and can go unabated due to limited or no regulations. This is partly due to little or no data available⁷. In the IGP region, the sources of air pollution include industrial activities⁸, vehicular emissions⁹, burning of agricultural crop residue¹⁰, and transfer of trans-boundary pollutants. In particular, trans-boundary sources of air pollutants can impact air quality across countries as has been observed in the IGP with India being the major source of pollution affecting both Pakistan and Bangladesh. The latter is particularly vulnerable, where most air masses from the Indian IGP region transports pollutants across borders¹¹ contributing up to 40% of pollution in Bangladesh-India bordering cities¹².

In South Asia, the brick kiln industry contributes not just to greenhouse gas emissions (GHG) but also to non-GHG emissions. Asian continent contributes to 87% of the total 1.5 trillion clay bricks manufactured annually with South-Asian countries (Pakistan, India, Nepal, and Bangladesh) producing 20% of those second only to China¹³. The kilns in South Asia are fueled by low- to medium-grade coal. Non-GHG emissions associated with producing bricks include particulate matter (PM), sulphur dioxide (SO₂), nitrogen oxides (NO_x), and carbon monoxide (CO) emissions pose a health hazard. A World Bank report claims that the brick kiln industry can be responsible for up to 91% of air pollution in some cities of the three South Asian countries mentioned¹⁴.

The short- and long-term exposure to the listed pollutants poses a health risk. Fine particulate matter such as PM_{2.5} or smaller can penetrate deep into the lungs, circulatory system, and other organs¹⁵. Exposure to PM_{2.5} can cause cardiovascular, respiratory, and pulmonary diseases with the potential of being fatal¹⁶. Research has also shown evidence of air pollution leading to epigenetic modifications which can be inherited across generations¹⁷. This highlights the potential impacts of air pollution on the neurodevelopmental growth of future generations.

¹School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, 46000, Pakistan. ²Department of Geoinformatics - Z_GIS, University of Salzburg, 5020, Salzburg, Austria. ³Smith School of Enterprise and Environment, University of Oxford, South Parks Rd, Oxford OX1 3QY, England, UK. e-mail: hassan.sheikh@smithschool.ox.ac.uk

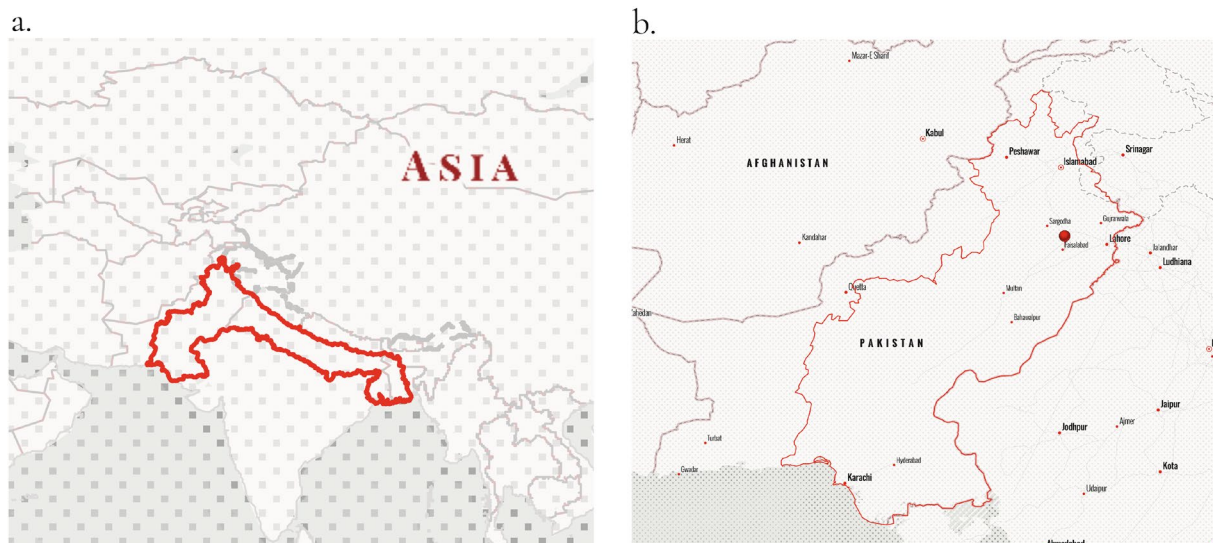


Fig. 1 Study Area Map - **(a)** The Indian-Gangetic Plain (IGP) region, spanning across Bangladesh, India, and Pakistan **(b)** Focuses on the IGP region within Pakistan, which is the specific area of interest for kiln detection in this study.

Studies have shown a link between PM, CO, NO₂, and SO₂ exposure during pregnancy and an increased probability of premature births and low birth weights in infants^{18,19}. This is a critical issue, given Pakistan has one of the highest fertility rates in South Asia²⁰. Similarly, prolonged exposure to air pollution significantly reduces the life expectancy, such as 4.3 years in Pakistan²¹, 5.3 years in India by²², and a staggering 6.8 years in Bangladesh²³ on average.

Despite knowing the health hazards of air pollution, a major challenge remains: accurately monitoring the impacts of various pollutants, identifying and understanding their sources.

The lack of a publicly available dataset for brick kilns in Pakistan also correlates with the informal and often illegal operation of the facility. The kilns in South Asia are outdated but research suggests that the substitution of these traditional brick kiln practices such as Bull's Trench Kiln (BTK) and Fixed Chimney Bull's Trench Kiln (FCBK) adopted in the South-Asia's IGP region with more efficient technologies such as ZigZag or Hoffman Kilns may reduce CO and PM emissions by over 60%^{13,24}. Despite China being the world's largest producer of brick kilns, its manufacturing is dominated by modern technologies using Hoffman Kilns, allowing it to reduce emissions of toxic pollutants and greenhouse gas emissions²⁵. Therefore, by determining the location of these polluting assets, we can estimate pollution exposure to nearby population and demand for better regulations.

Previously, there have been efforts to detect brick kilns at scale using satellite imagery in Bangladesh²⁶ and India²⁷. An emission inventory developed for the Indian region of IGP has declared brick kilns to be the top-most priority for assessing particulate matter emissions with Uttar Pradesh, India having the highest number of reported brick kilns²⁸. Previous works have used field surveys and remote sensing to evaluate energy consumption and enumerate brick kilns in India²⁹. Deep Neural Networks (DNNs) have been used to demonstrate the use of high-resolution satellite imagery for the West Bengal region in India²⁷; whereas, another study highlights transfer learning and self-supervised learning techniques for detecting 7477 brick kilns in five states of India with an accuracy of 81.72%³⁰. For Bangladesh, high-resolution satellite imagery and a deep learning backed approach have been used for the identification of brick kilns, reporting 94.2% accuracy with 88.7% precision²⁶.

In Pakistan, despite efforts to use satellite imagery and AI to detect brick kilns in certain regions of the country^{31,32}, there remains a data gap in large-scale satellite monitoring of brick kilns.

To this end, we have developed an open source asset-level dataset of brick kilns in Pakistan's IGP region. The final dataset differentiates between Fixed Chimney Bull's Trench Kiln (FCBK) and ZigZag kiln types. The dataset contains detailed brick kiln asset characteristics including: id (unique identifier for each brick kiln site), location specific information (latitude, longitude, administrative state or region of the kiln), type (classification of brick kiln (FCBK or ZigZag), and number of schools, hospitals, estimated populations within a 1 km radius.

Methods

We collected open-source satellite imagery to develop an asset-level brick kiln dataset. The Pakistani portion of the IGP primarily consists of the vast Indus River plain, covering approximately 200,000 square miles (518,000 square kilometers) of fertile land³³. This region encompasses a significant part of Pakistan's agricultural heartland, including most of Punjab province, as well as part of Sindh and Khyber Pakhtunkhwa (KP) province.

Although brick kilns are easily visible in satellite images, precisely identifying their locations is time-consuming and expensive. To address this issue, we utilize a combination of openly available low-resolution

satellite imagery and commercial high-resolution satellite imagery that helps to reduce costs and speed up the identification process.

The methodology to build the dataset is divided into two phases to effectively identify and classify brick kilns in satellite imagery. In the first phase, we applied a Random Forest classifier to low-resolution RGB Sentinel-2 imagery for initial pixel-wise detection followed by a post-processing step across the entire region shown in Fig. 1(b). This allowed for extensive coverage but was hindered by the similarity in color profiles between brick kilns and surrounding areas, leading to occasional misclassifications. To overcome this, we integrated a YOLOv8 object detection model, using high-resolution imagery from the Google Maps Static API³⁴ to differentiate between FCBK and ZigZag kilns and validate points identified by the Random Forest model. This high-resolution pipeline was also utilized in regions without Sentinel-2 data. By primarily relying on open-source low-resolution imagery and selectively using high-resolution imagery when necessary, we created a comprehensive dataset of brick kiln locations across the study area, addressing the existing data gap in Pakistan.

Phase One: Random Forest for Pixel-Wise Identification. In this phase, we leveraged Random Forest to perform an initial, pixel-wise classification of brick kilns using Sentinel-2 imagery followed by a dedicated post-processing pipeline to refine detections, reduce false positives, and enhance the accuracy of the identified kiln locations.

Data Collection and Annotation for Random Forest. The data collection process focused on acquiring satellite imagery for the study area in the Pakistani IGP region. We utilized Sentinel-2 imagery from July 1, 2023, to July 15, 2023, a period when brick kilns are typically active in Pakistan. Specifically, we used Sentinel-2 MSI (MultiSpectral Instrument) Level-2A imagery, which provides surface reflectance data for precise detection and analysis³⁵. This imagery was accessed using the Google Earth Engine platform (<https://earthengine.google.com>) via the dataset `COPERNICUS/S2_SR`, which is openly available under the Copernicus Open Access License.

The dataset included the RGB bands of the Sentinel-2 imagery and served as our primary dataset for brick kiln detection across most of the study area. To ensure high-quality imagery, we applied a cloud cover removal process and downloaded only images with less than 2% cloud cover. The cloud masking process involves identifying and removing pixels affected by clouds and cirrus from the imagery. This is achieved by evaluating quality assessment flags that indicate the presence of these atmospheric conditions. By applying this mask, we ensure that only clear and cloud-free data is utilized. The Pakistani IGP region as shown in Fig. 1(b), was divided into measuring $100 \times 100 \text{ km}^2$ grids, totaling approximately 60 grids. To effectively train the models, we annotated satellite images of each grid tile with different land cover classes. This process involved manually labeling the images to categorize different land cover types, such as urban areas, vegetation, water bodies, bare land, and brick kilns as the main target feature. Careful annotation was crucial to avoid confusion in the model training process and to ensure the accurate classification of diverse land cover types within the area of interest. The visual appearances of these land cover types are illustrated in Fig. 2 and their semantic characteristics are delineated in Table 1. By annotating all 60 grids across the entire region rather than focusing on a smaller subset, this approach is estimated to have reduced false positives by at least threefold. We categorized the annotations into ten classes as shown in Table 3.

Training Setup and Pixel-Wise Classification. For our classification task, we employed a Random Forest classifier, a robust ensemble learning method known for its ability to handle diverse datasets. This model was selected due to its proven efficacy in managing high-dimensional data, such as the Sentinel-2 imagery, and its capability to accommodate both continuous and categorical variables³⁶. The dataset was divided into 80% training and 20% testing, ensuring that the model was trained on a substantial portion while reserving enough data for evaluation. The Random Forest model was trained with 500 decision trees (estimators), each constructed using a random subset of the training data and features to mitigate over-fitting and enhance the model's generalisability. At each split, a maximum of ten features were considered, with a maximum tree depth of 50, a minimum of two samples per split, and a minimum of one sample per leaf. This configuration promoted efficient feature selection and optimized model performance on the test data.

The model achieved a high recall of 97% and a precision of 72% on the test dataset for detecting brick kilns. A recall rate of 97% indicates that the model successfully identified 97% of the actual brick kiln pixels in the dataset, meaning that it captured nearly all instances of the brick kilns, with very few false negatives. However, the precision rate of 72% shows that, among the instances labeled as brick kilns by the model, only 72% were actually brick kilns, indicating a considerable number of false positives. In this context, a false positive occurs when the model incorrectly labels a non-brick kiln feature as a brick kiln. High recall with low precision often implies that while the model is very sensitive in detecting potential brick kilns, it lacks specificity, and mistakenly identifies other structures or features with similar visual characteristics as brick kilns. These false positives can lead to overestimations in the count or area of brick kilns, which would skew the analysis and limit the practical utility of the results. To address this imbalance and improve the accuracy of our brick kiln detection, we recognized the need for a post-processing pipeline.

Sentinel-2 Grid Selection For Inferencing. Initially, we used $5 \times 5 \text{ km}$ grids to cover large areas, but this grid size proved inadequate for detecting smaller, dispersed kilns due to Sentinel-2's 10m resolution. As shown in Supplementary Fig. 1, where (a) represents a $5 \times 5 \text{ km}$ grid and (b) represents a $1 \times 1 \text{ km}$ grid, the larger grid size makes it challenging to visually identify individual brick kilns. The $5 \times 5 \text{ km}$ grid limited the effectiveness of post-processing steps, resulting in reduced detection accuracy. Switching to $1 \times 1 \text{ km}$ grids allowed for more focused analysis, better aligned with Sentinel-2's resolution, and improved the precision of our post-processing steps. This adjustment reduced false positives and allowed us to capture smaller kilns more

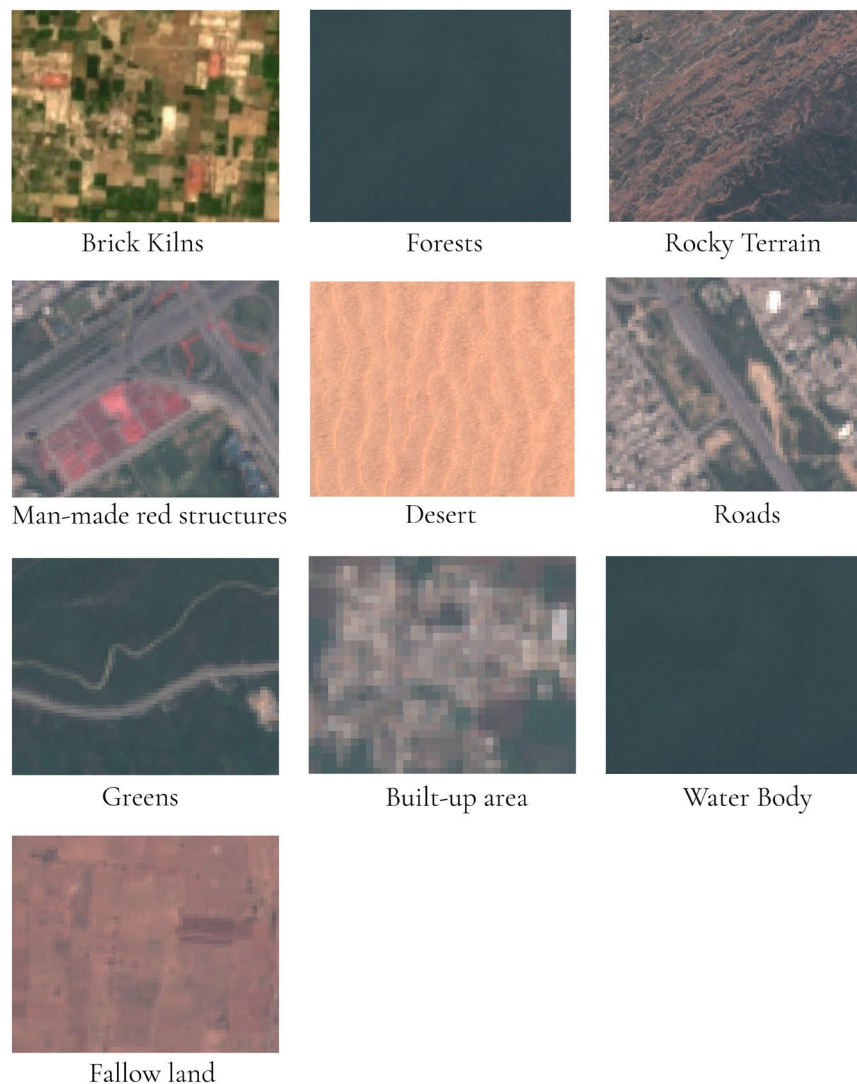


Fig. 2 Image tiles illustrating different classes labelled for training our Random Forest Classifier.

accurately. Consequently, the 1×1 km grid proved to be the optimal choice for accurate brick kiln identification across the study area.

Post Processing and Geolocating. After training the model, we applied the Random Forest classifier to Sentinel-2 imagery across the entire IGP region, where each Sentinel-2 image covers an area of 1×1 km². Following the initial pixel-wise classification using Random Forest across the entire IGP region as illustrated in Fig. 3(a), we applied a post-processing step to enhance the accuracy of brick kiln detection. This step was crucial for reducing noise and improving the coherence of detected regions. The post-processing pipeline, shown in Fig. 3(b), follows a structured sequence of steps. The process begins by generating a binary mask from the Random Forest model's pixel-wise classification results, which identifies areas that are likely to contain brick kilns. This results in a new image consisting of two classes: detected brick kilns, shown as red pixels, and the background, represented in black. Isolated pixels are removed to reduce false positives. Morphological closing is applied to eliminate noise and consolidate fragmented areas, forming coherent clusters. These red pixels are clustered based on proximity, with each cluster corresponding to a distinct kiln. The geometric center of each cluster is computed to determine precise kiln locations, converted into geographical coordinates, and redundant detections within a 20 meter radius are eliminated. Finally, the number of centers is capped at fifteen per square kilometer, as it is highly improbable for more than fifteen brick kilns to exist within such a confined area, ensuring a significant reduction in false positives.

Sentinel-2 Pipeline Results. The machine learning pipeline was evaluated across four regions: Khyber Pakhtunkhwa (KP), northern Punjab, southern Punjab, and Sindh. The entire IGP region comprises 20,873 data points, with 11,536 points from southern Punjab and Sindh, and 9,337 points from Northern Punjab and KP. The model performed particularly well in Northern Punjab and KP, as the detected points closely matched the

T0	T1	T2	Semantic Characteristic
Land Cover	Vegetated Areas	Green Areas	Areas covered with grass and maintained vegetation with tones of green
		Forests	Large areas with continuous dense, dark green patches with dominance of trees
	Non-Vegetated Areas	Fallow Lands	Light brown, reddish tones with large, open rectangular patches with visible soil and sparse or no vegetation
		Desert	Uniform land with a sandy texture, little vegetation, brown or beige tones, and ripple like patterns expanding large areas with little variation in landscape
		Rocky Terrain	Rough, jagged patterns with reddish brown tones and uneven textures with pronounced vertical features
	Water Bodies		Rivers, lakes, ponds, and other standing or flowing bodies of water
	Urbanized Structures	Urban Areas	Pixelated areas with mixtures of colors appear as dense and complex patches of structures
		Redroof Structures	Red human made structures with rectangular or geometric patterns with visible boundaries
		Roads	Long, straight, or curved linear features typically in gray tones sometimes with intersections
	Industrial Use	Brick Kilns	Ovular or rectangular structures surrounded either by vegetation or barren land with a distinguishable reddish-brown tone.

Table 1. Categorisation of Land Cover Types with Semantic Characteristics for Each Sub-Class.

Region	Total Points Identified by RF	Brick Kilns Detected by YOLO	Percentage Detected (%)
Northern Punjab and KP	9,337	6,706	71.8%
Sindh and Southern Punjab	11,536	4,271	37.0%
Regions without Sentinel-2 Imagery (Google Maps Static API only)	—	301	—
Total	20,873	11,277	54.0%

Table 2. Performance of Random Forest Model in Identifying Brick Kiln Points Across Different Regions and Subsequent YOLO Detection Results.

Index	Class	Number of Samples (Pixels)
1	Brick Kilns	25,678
2	Redroof Structures	950
3	Water Bodies	6,578
4	Green Areas	13,234
5	Forests	4,300
6	Fallow Lands	21,893
7	Desert	6,918
8	Urban Areas	5,435
9	Roads	3,462
10	Rocky Terrain	8,214

Table 3. Annotated Pixel Counts for Land Cover Classes in Sentinel-2 Imagery for Random Forest Classifier Training.

expected number of brick kilns in these areas. However, in Southern Punjab and Sindh, the model produced a higher rate of false positives. Despite this, the post-processing pipeline significantly improved results, reducing false positives by twenty-fold. Without this pipeline, we estimated that over 400,000 points would have been generated, making cross-verification prohibitively costly. There were also areas where Sentinel-2 data was unavailable and the model could not detect any brick kilns in those regions.

Advancing to YOLO: Building on Random Forest Findings. In the random forest phase, detecting brick kilns in the Sindh and Southern Punjab regions presented a challenge—kilns and surrounding areas exhibiting red hue. This resemblance led to high false positive rates. Since the random forest algorithm analyzes each pixel independently, it loses important spatial context, making it difficult to distinguish between neighboring objects. Additionally, though fewer, false positives also occurred in Northern Punjab and KP. Given these challenges and the need to classify brick kilns into FCBK and ZigZag types, we transitioned to YOLO, which improves spatial awareness by dividing the image into grids, predicting bounding boxes, and assigning class probabilities, resulting in more accurate object detection. YOLO's deep learning architecture autonomously learns relevant features during the training process, eliminating the need for manual feature selection required in methods such

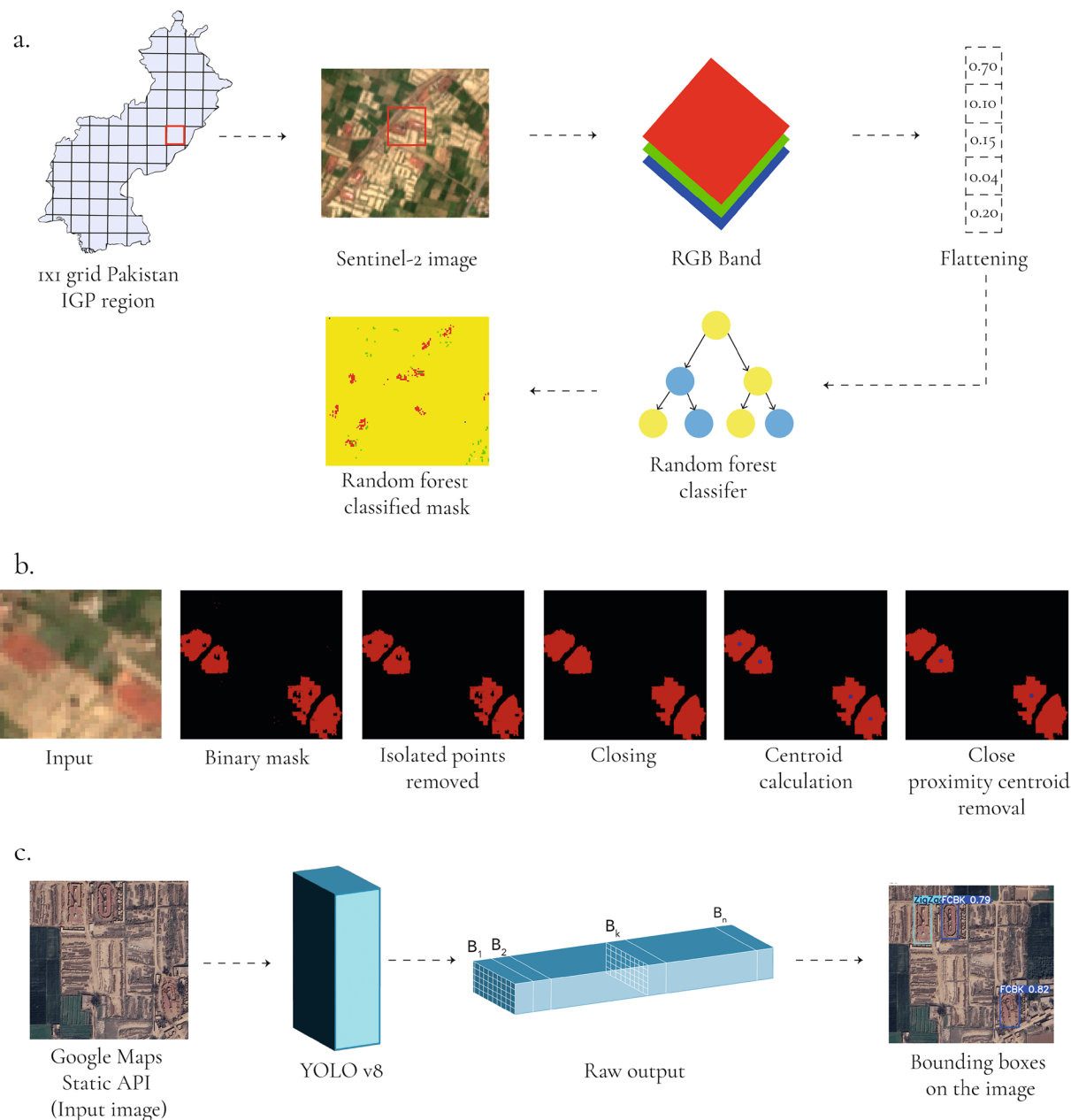


Fig. 3 (a) The figure outlines the process of brick kiln detection in the Indus-Gangetic Plain using Sentinel-2 imagery. A 1×1 grid of the region is extracted, RGB bands are flattened into feature vectors, and a Random Forest classifier generates a mask identifying brick kiln locations. (b) The figure shows the visual representation of the steps involved in post-processing pipeline to accurately geolocate brick kilns in the image. (c) The figure illustrates the process of using the YOLOv8 model to detect brick kilns from Google Maps imagery, resulting in bounding boxes on the input image.

as Random Forest. Our kiln type identification pipeline was ran on two types of regions: (a) 20,873 locations identified by our Random Forest model and (b) the region where Sentinel-2 imagery was unavailable (Supplement Figure S2).

Phase Two: YOLO for High-Resolution Imagery. In this phase of the pipeline, we used high-resolution satellite imagery and YOLO to enhance the detection of brick kiln points identified in low-resolution imagery by the Random Forest classifier.

Data Collection and Annotation for YOLO. We extracted high-resolution imagery through the Google Maps Static API Map data ©2024 Google. (<https://developers.google.com/maps/documentation/maps-static>), using a zoom level of 17 and a scale of 2, which produced images with dimensions of 1280×1280 pixels. The scale factor effectively doubled the resolution from the default 640×640 pixels, providing a greater level of detail. For these

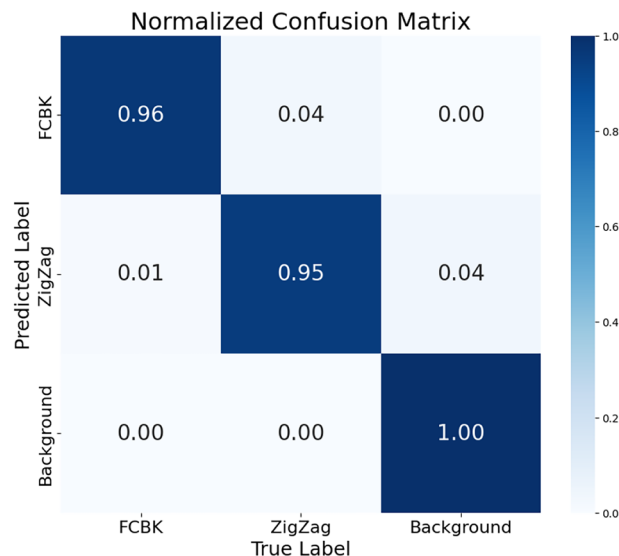


Fig. 4 Normalised Confusion Matrix for YOLOv8n.

high-resolution images, the annotation process was straightforward. We annotated approximately 375 FCBK brick kilns and 295 ZigZag brick kilns using bounding boxes. While Oriented Bounding Boxes (OBBs) generally provide a more precise representation of an object's shape and orientation³⁷, we opted for standard bounding boxes to streamline geolocation. Since our objective was to obtain only a single central coordinate per kiln, the added computational expense of OBBs was unnecessary, making regular bounding boxes a more efficient choice.

Training Setup. The YOLOv8n³⁸ model was trained using a dataset split into 80% for training, 10% for validation, and 10% for testing, ensuring robust evaluation and fine-tuning throughout the process. The model was trained for 250 epochs, with a batch size of 8 and an image resolution set to 1280×1280 . Early stopping was applied with patience of 100 epochs, and the model checkpoint from epoch 114 was selected due to no further performance improvement after this point. The training utilized an initial learning rate (lr0) of 0.01 with a learning rate decay factor (lrf) of 0.01. The optimizer was set to auto, with a momentum value of 0.937 and a weight decay of 0.0005. A warmup period of 3 epochs was employed, with a warmup momentum of 0.8 and a warmup bias learning rate of 0.1. Data augmentation techniques included RandAugment with a probability of erasing set to 0.4, translation at 0.1, and scaling at 0.5. The mosaic augmentation was activated with a factor of 1.0, and horizontal flipping was applied with a probability of 0.5. Other transformations, such as rotation (degrees) and shear, were kept at 0.0. Non-maximum suppression (NMS) was configured with an IoU threshold of 0.7, and the maximum number of detections per image was capped at 10. The model used overlap masks with a ratio of 4 and was trained with 8 workers. Automatic Mixed Precision (AMP) was enabled to optimize computational efficiency.

The YOLO object detection model was trained using High-Resolution satellite imagery. The training yielded strong results, with a mean Average Precision (mAP) at 50% Intersection over Union (IoU) reaching 95%, and mAP at 95% IoU reaching 52%. The normalised confusion matrix in Fig. 4 illustrates the model's classification performance across the three classes on the test dataset: FCBK, ZigZag, and background. A high-level view of obtaining class and bounding box predictions is shown in Fig. 3(c).

Inference. For the YOLO pipeline, inferencing was conducted on two key areas: the 20,873 points identified by our Random Forest model and regions where Sentinel-2 data was unavailable. Among the 20,873 identified points, some were located in close proximity, raising the risk of overlapping brick kiln detections. To mitigate this, we used a zoom level of 17, where each image covers an area of 0.45 km². We grouped coordinates within a 0.45 km² radius, downloading a single image per group to minimize redundancy and optimize data management. This grouping approach effectively reduced double counting and minimized the volume of imagery required, while addressing the high false positive rate observed in this region. Additionally, if a kiln appeared split across two contiguous images as shown in Fig. 5(a),(b), only one instance was counted by removing points within a 12-meter radius of each other. To accurately calculate the geographic coordinates for each detected bounding box, we applied a method described in the Supplementary Section I, which maps bounding box centers to geographic coordinates (Equations 1–6). This approach ensures precise localization of brick kilns within the satellite imagery.

For areas lacking Sentinel-2 coverage, high-resolution imagery was obtained for the entire region to ensure comprehensive detection of brick kilns. After the completion of the second phase of the entire pipeline, we detected more than 11,000 brick kilns in the region.

The annotations for both our Random Forest and YOLO models were performed by experts familiar with satellite imagery and brick kiln detection. To ensure consistency and reduce annotation errors, the annotations were cross-validated by different team members. This process helped establish a strong ground truth dataset for training and validating our models. Our validation process was designed in two parts: before running inference

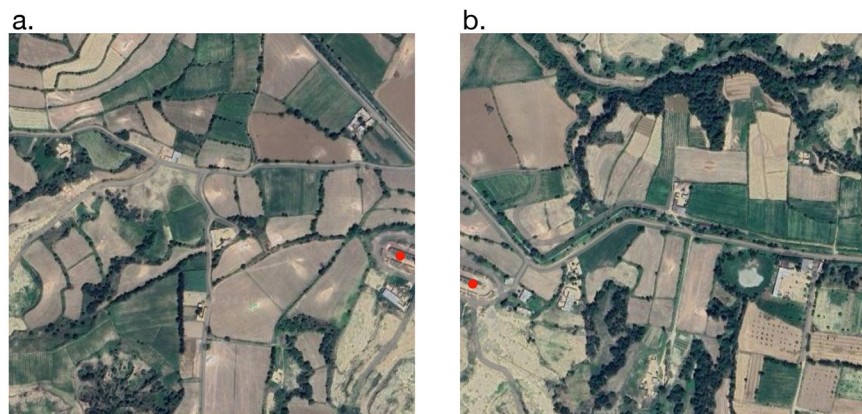


Fig. 5 (a) and (b) represent two adjacent image tiles where the same kiln is partially visible in both.

Region	Recall	Precision
Southern Punjab	0.90	0.65
Sindh	0.88	0.60
Northern Punjab	0.99	0.78
KP	0.95	0.74

Table 4. Region-wise Precision and Recall for Brick Kilns on Test Dataset Using Random Forest Classifier.

over the entire region and after the full inference was completed. Prior to the large-scale inference, we evaluated the performance of our models using precision, recall, and F1 scores across different regions. This pre-inference validation was critical for assessing whether the model was adequately trained and performing as expected. As the Table 4 indicates, we observed a high number of false positives from the Random Forest Classifier which necessitated further refinement through the post-processing pipeline outlined in the paper. The final results, after running both the low-resolution (Sentinel-2 and Random Forest) and high-resolution (YOLOv8) pipelines, were manually verified by experts to ensure the accuracy of detected kiln locations.

Downstream analysis. The proximity metrics (number of schools, hospitals, and population) were computed using spatial overlay analysis with publicly available datasets. Amenities data were sourced from OpenStreetMap (OSM), while gridded population estimates were obtained from the [LandScan Global 2023 dataset](#). A 1 km circular buffer was used around each kiln to calculate proximity metrics.

For emission estimates, the data record includes pollutants such as PM_{10} , $PM_{2.5}$, SO_x , and NO_x for each brick kiln site. These emissions were calculated using a bottom-up approach based on production data and standardised emission factors, accounting for operational days and the types of fuel used during kiln operations. See Supplementary Section II for more details. It is expressed in kilograms per day, providing insight into potential daily pollutant emissions for each site. These measures are helpful in the assessment of the number of people potentially impacted by emissions and can be crucial for developing targeted regulatory or mitigation strategies. Additionally, each kiln in the dataset is categorized by its operational type: FCBK and ZigZag as shown in the Fig. 6(b). This classification is essential as different kiln types have varying fuel efficiencies and emission profiles. Zigzag kilns, for example, are generally more fuel-efficient and emit fewer pollutants than traditional kilns. This attribute enables comparative analysis across kiln types to assess the environmental impact and will assist regulatory bodies in identifying kilns that may benefit from technological upgrades.

Data Records

The dataset⁴¹, available at [Zenodo Data Records](#) (reference: 14038648), is provided in three formats: CSV, shapefile (.shp), and GeoJSON (.geojson), facilitating spatial analysis workflows. It contains geospatial and contextual attributes for each brick kiln site, including location coordinates (WGS84 CRS), emission estimates (g/kg of bricks produced), kiln type, and proximity to sensitive sites such as schools, hospitals, and populated areas (Table 5). The dataset⁴¹ captures critical information related to brick kiln emissions and associated risks to nearby populations, with detailed profiling by country, district, and proximity analysis within a 1 km radius (Fig. 6a,c). For clarity, a sample data entry along with descriptions of each column is provided in Table 5.

Technical Validation

To assess the completeness and reliability of our dataset, we compared it against reported brick kiln numbers. The estimated reported kilns in the ‘brick belt’ of India, Pakistan, and Bangladesh is 55,387; out of which around 11,000 are in Pakistan³⁹.

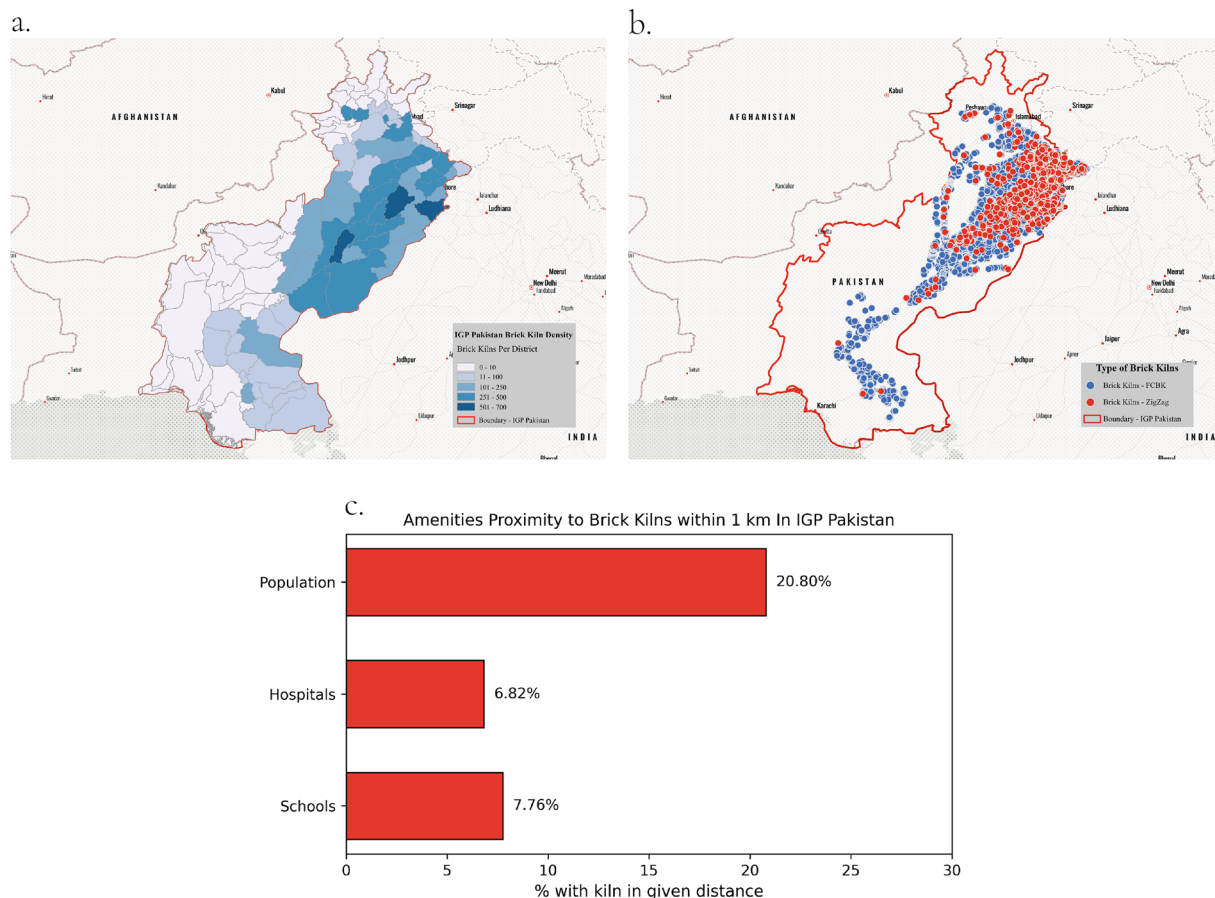


Fig. 6 Overview of brick kiln distribution, types, and proximity to sensitive areas within IGP-Pakistan. **(a)** Distribution of brick kiln density across IGP-Pakistan, highlighting areas with high kiln concentrations and potential pollution hotspots. **(b)** Classification of kiln types (FCBK and zigzag) showing regional differences in kiln technology. **(c)** Percentage of schools, hospitals, and population density within 1 km of kilns, indicating exposure risks for nearby communities.

Column	Example Value	Description
id	BKP-1	Unique identifier for the kiln site
lat	26.133092	Latitude (decimal degrees)
lon	68.565789	Longitude (decimal degrees)
state	Sindh	Administrative region of kiln
type	FCBK	Kiln classification (ZigZag or FCBK)
schools1km	0	Schools within 1 km radius
hosplkm	0	Hospitals within 1 km radius
pop1km	1274	Estimated population within 1 km

Table 5. Example row from the dataset with variable descriptions.

As outlined in Table 2, the trained YOLO model successfully detected approximately 11,277 brick kilns. Of these, 6,706 were located in Northern Punjab and Khyber Pakhtunkhwa (KP), while 4,271 were detected in Sindh and Southern Punjab. These findings align with our initial hypothesis regarding the performance of the Random Forest classifier. In Northern Punjab and KP, where the Random Forest Classifier identified 9,337 points, 6,706 were confirmed to be brick kilns, indicating strong performance in these regions. Conversely, the model showed moderate performance in Sindh and Southern Punjab, with 4,271 brick kilns detected out of 11,536 points. Due to the effectiveness of the initial post-processing pipeline explained earlier, we downloaded a total of 60,000 high-resolution images for analysis including the regions where Sentinel-2 data was unavailable reducing the costs associated with the high-resolution imagery.

For emissions estimates, we validated by comparing our bottom-up approach results with previous studies on brick kiln emissions in similar regions⁴⁰. The proximity of schools, hospitals, and populated areas within 1 km

of each kiln was assessed using the kiln points identified in our findings, combined with OpenStreetMap (OSM) data and (GIS) techniques.

Usage Notes

Users can access and download these files through our [Zenodo Data Records](#)⁴¹ for integration into GIS and data analysis software, such as QGIS, ArcGIS, or R/Python. For further details on attribute descriptions, units, and abbreviations, please refer to the supplementary document S1. Researchers can utilize the *csv*, *shapefile*, or *geojson* as they prefer, to analyze spatial distributions, pollutant estimates, and proximity risks. Users can conduct spatial queries, overlay additional environmental layers, and perform risk assessment modeling. For accurate and up-to-date demographic data, users may cross-reference with local census or updated OSM data, especially for studies involving dynamic population or infrastructure changes.

Code availability

The code notebook including model weights for production of this dataset are open-source and available at the APAD repository on [Github](#). For the low-resolution pipeline, we used Google Earth Engine's Python-based API to download GeoTIFF files from Sentinel-2 imagery. Rasterio was employed for geospatial data handling, Scikit-learn library in Python was used to train the Random Forest Classifier, and Python OpenCV was used for post-processing steps such as noise removal and clustering of detected brick kilns. For the high-resolution pipeline, we used Ultralytics' Python API to handle the training and inferencing of the YOLOv8 model. High-resolution imagery from Google Maps Static API was processed to improve detection accuracy.

Received: 21 November 2024; Accepted: 7 May 2025;

Published online: 20 May 2025

References

1. Bank, W. *Striving for Clean Air: Air Pollution and Public Health in South Asia*. South Asia Development Matters (World Bank, Washington, DC, 2023).
2. IQAir. *World's most polluted cities 2023*. Accessed: 2023-12-09 (2023).
3. Shi, Y., Bilal, M., Ho, H. C. & Omar, A. Urbanization and regional air pollution across south asian developing countries: A nationwide land use regression for ambient PM_{2.5} assessment in pakistan. *Environmental Pollution* **266**, 115145, <https://doi.org/10.1016/j.envpol.2020.115145> (2020).
4. The Energy and Resources Institute (TERI). *Scoping study for south asia air pollution*. Commissioned by the South Asia Research Hub, Department for International Development, Government of UK (2019).
5. A critical review of managing air pollution through airshed approach. **9**.
6. Rahman, M. M. *et al.* Comprehensive evaluation of spatial distribution and temporal trend of no, so and aod using satellite observations over south and east asia from 2011 to 2021. *Remote Sensing* **15**, <https://doi.org/10.3390/rs15205069> (2023).
7. Mazhar, U. & Elgin, C. Environmental regulation, pollution and the informal economy. *SBP Research Bulletin* **9**, 62–81 Accessed on: 2024-10-10 (2013).
8. Sumaira & Siddique, H. M. A. Industrialization, energy consumption, and environmental pollution: Evidence from south asia. *Environmental Science and Pollution Research* **30**, 4094–4102, <https://doi.org/10.1007/s11356-022-22317-0> (2022).
9. Abdul Jabbar, S. *et al.* Air quality, pollution and sustainability trends in south asia: A population-based study. *International Journal of Environmental Research and Public Health* **19**, <https://doi.org/10.3390/ijerph19127534> (2022).
10. Ravindra, K. *et al.* Appraisal of regional haze event and its relationship with pm_{2.5} concentration, crop residue burning and meteorology in chandigarh, india. *Chemosphere* **273**, 128562, <https://doi.org/10.1016/j.chemosphere.2020.128562> (2021).
11. Rana, M. M., Mahmud, M., Khan, M. H., Sivertsen, B. & Sulaiman, N. Investigating incursion of transboundary pollution into the atmosphere of dhaka, bangladesh. *Advances in Meteorology* **2016**, 1–11, <https://doi.org/10.1155/2016/8318453> (2016).
12. Dibya, T. B., Prama, A. Y. & Dewan, S. M. R. Poor respiratory health is a consequence of dhaka's polluted air: A bangladeshi perspective. *Environmental Health Insights* **17**, 1–4, <https://doi.org/10.1177/11786302231206126> (2023).
13. Seay, B., Adetona, A., Sadoff, N., Sarofim, M. C. & Kolian, M. Impact of south asian brick kiln emission mitigation strategies on select pollutants and near-term arctic temperature responses. *Environmental Research Communications* **3**, 061004, <https://doi.org/10.1088/2515-7620/ac0a66> (2021).
14. Eil, A., Li, J., Baral, P. & Saikawa, E. Dirty stacks, high stakes: An overview of brick sector in south asia. World Bank Document (2020).
15. Pryor, J. T., Cowley, L. O. & Simonds, S. E. The physiological effects of air pollution: Particulate matter, physiology and disease. *Frontiers in Public Health* **10**, <https://doi.org/10.3389/fpubh.2022.882569> (2022).
16. Guo, J. *et al.* Long-term exposure to particulate matter on cardiovascular and respiratory diseases in low- and middle-income countries: A systematic review and meta-analysis. *Frontiers in Public Health* **11**, 1134341, <https://doi.org/10.3389/fpubh.2023.1134341> (2023).
17. Shukla, A. *et al.* Air pollution associated epigenetic modifications: Transgenerational inheritance and underlying molecular mechanisms. *Science of The Total Environment* **656**, 760–777, <https://doi.org/10.1016/j.scitotenv.2018.11.381> (2019).
18. Bazyar, J. *et al.* A comprehensive evaluation of the association between ambient air pollution and adverse health outcomes of major organ systems: a systematic review with a worldwide approach. *Environmental Science and Pollution Research* **26**, 12648–12661, <https://doi.org/10.1007/s11356-019-04874-z> (2019).
19. World Health Organization (WHO). *Health impacts - types of pollutants* (2023). Accessed: 2023-12-09.
20. World Bank. *Fertility rate, total (births per woman) - pakistan* (2022). Accessed: 2023-12-09.
21. Greenstone, M. & Fan, Q. C. *Pakistan fact sheet* (2023). AIR QUALITY LIFE INDEX[®] (AQLI).
22. Greenstone, M. & Fan, Q. C. *India fact sheet 2023* (2023). AIR QUALITY LIFE INDEX[®] (AQLI).
23. Greenstone, M. & Fan, Q. C. *Bangladesh fact sheet 2023* (2023). AIR QUALITY LIFE INDEX[®] (AQLI).
24. Kumar, N. & Chaurasia, A. Climate change and its impact on south asian agriculture. In *Sustainable Agriculture in the Era of Climate Change*, 879–890, https://doi.org/10.1007/978-3-030-73943-0_52 (Springer, 2021).
25. Bashir, Z. *et al.* Investigating the impact of shifting the brick kiln industry from conventional to zigzag technology for a sustainable environment. *Sustainability* **15**, <https://doi.org/10.3390/su15108291> (2023).
26. Lee, J. *et al.* Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences (PNAS)* **118**, e2018863118, <https://doi.org/10.1073/pnas.2018863118> (2021).
27. Paul, A., Bandyopadhyay, S. & Raj, U. Brick kiln detection in remote sensing imagery using deep neural network and change analysis. *Spatial Information Research* **30**, 607–616, <https://doi.org/10.1007/s41324-022-00458-1> (2022).

28. Ghosh, A. *et al.* A district-level emission inventory of anthropogenic pm_{2.5} from the primary sources over the indian indo gangetic plain: Identification of the emission hotspots. *Science of The Total Environment* **914**, 169865, <https://doi.org/10.1016/j.scitotenv.2023.169865> (2024).
29. Tibrewal, K. *et al.* Reconciliation of energy use disparities in brick production in india. *Nature Sustainability* **6**, 1248–1257, <https://doi.org/10.1038/s41893-023-01165-x> (2023).
30. Mondal, R., Patel, Z. B., Jani, V. & Batra, N. Scalable methods for brick kiln detection and compliance monitoring from satellite imagery: A deployment case study in india. arXiv preprint ArXiv:2402.13796v1 (2024).
31. Tahir, R. *et al.* Brick kiln detection and localization using deep learning techniques. In *2021 International Conference on Artificial Intelligence (ICAI)*, 37–43, <https://doi.org/10.1109/ICAI52203.2021.9445267> (2021).
32. Nazir, U., Taj, M., Uppal, M. & Khalid, S. Mitigating climate and health impact of small-scale kiln industry using multi-spectral classifier and deep learning (2023).
33. Burki, S. J. & Ziring, L. Pakistan (2024). Encyclopedia Britannica, Accessed 17 September 2024.
34. Google Developers. Google maps static api documentation. <https://developers.google.com/maps/documentation/maps-static>. Accessed: 2025-04-23 (2024).
35. Sentinel-2 - missions - sentinel online - sentinel online. Available at: <https://sentinel.esa.int/web/sentinel/missions/sentinel-2> (2023).
36. Truong, X. Q. *et al.* Random forest analysis of land use and land cover change using sentinel-2 data in van yen, yen bai province, vietnam. In Nguyen, L. Q., Bui, L. K., Bui, X.-N. & Tran, H. T. (eds.) *Advances in Geospatial Technology in Mining and Earth Sciences*, 429–445 (Springer International Publishing, Cham, 2023).
37. Yi, J. *et al.* Oriented object detection in aerial images with box boundary-aware vectors. 2149–2158, <https://doi.org/10.1109/WACV48630.2021.00220> (2021).
38. Jocher, G., Qiu, J. & Chaurasia, A. Ultralytics YOLO (2023).
39. Boyd, D. S. *et al.* Slavery from space: Demonstrating the role for satellite remote sensing in mapping modern slavery. *Computers, Environment and Urban Systems* **68**, 76–88, <https://doi.org/10.1016/j.compenvurbysys.2017.09.007> (2018).
40. Abbas, A. *et al.* Assessment of long-term energy and environmental impacts of the cleaner technologies for brick production. *Energy Reports* **7**, 7157–7169, <https://doi.org/10.1016/j.egy.2021.10.072> (2021).
41. Hamdani, M. S. A., Zakir, K., Kushwaha, N., Fatima, S. E. & Sheikh, H. A. Brick kiln dataset for pakistan's igp region using ai, <https://doi.org/10.5281/zenodo.14038648> (2024).

Acknowledgements

We would like to thank Smith School of Enterprise and the Environment and Amazon Web Services (AWS) for funding and support respectively. We also appreciate AWS Cloud Credit for Research.

Author contributions

M.S.A.H. developed machine learning and deep learning pipelines. K.Z. worked on the remote sensing pipeline. M.S.A.H., K.Z., N.K. S.E.F., annotated the imagery. K.Z., M.S.A.H., S.E.F. validated the results. M.S.A.H., K.Z., S.E.F. H.A.S., N.K. wrote the first draft. H.A.S. and N.K. commented on the paper. H.A.S. conceptualised the need for a brick kiln dataset for Pakistan and supervised the project. All authors reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05148-9>.

Correspondence and requests for materials should be addressed to H.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025