remote sensing

MDPI

Article

From Prediction to Explanation: Using Explainable AI to Understand Satellite-Based Riot Forecasting Models

Scott Warnke [†,‡] **and Daniel Runfola** [*,‡]

Department of Applied Sciences, William & Mary, Williamsburg, VA 23185, USA; sdwarnke@wm.edu
* Correspondence: dsmillerrunfol@wm.edu
† Current address: 540 Landrum Drive, Williamsburg, VA 23185, USA.
‡ These authors contributed equally to this work.

Abstract: This study investigates the application of explainable AI (XAI) techniques to understand the deep learning models used for predicting urban conflict from satellite imagery. First, a ResNet18 convolutional neural network achieved 89% accuracy in distinguishing riot and non-riot urban areas. Using the Score-CAM technique, regions critical to the model's predictions were identified, and masking these areas caused a 20.9% drop in the classification accuracy, highlighting their importance. However, Score-CAM's ability to consistently localize key features was found to be limited, particularly in complex, multi-object urban environments. Analysis revealed minimal alignment between the model-identified features and traditional land use metrics, suggesting that deep learning captures unique patterns not represented in existing GIS datasets. These findings underscore the potential of deep learning to uncover previously unrecognized socio-spatial dynamics while revealing the need for improved interpretability methods. This work sets the stage for future research to enhance explainable AI techniques, bridging the gap between model performance and interpretability and advancing our understanding of urban conflict drivers.

Keywords: deep learning; convolutional neural networks; satellite imagery; explainable AI; land use/land cover

check for updates

Academic Editor: Zhenwei Shi

Received: 26 November 2024
Revised: 13 January 2025
Accepted: 15 January 2025
Published: 17 January 2025

Citation: Warnke, S.; Runfola, D. From Prediction to Explanation: Using Explainable AI to Understand Satellite-Based Riot Forecasting Models. *Remote Sens.* **2025**, *17*, 313. https://doi.org/10.3390/rs17020313

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Understanding why conflict arises, particularly in urban areas, is critical for addressing its causes and mitigating its impacts [1]. Deeper insights into urban conflict offer numerous benefits: city planners can better anticipate and manage human behavior in these spaces [2], while modeling riots and protests helps uncover the underlying dynamics driving these events [3,4]. In previous work [5], it has been shown that satellite imagery—in conjunction with deep learning approaches—could be leveraged to predict the most likely geospatial locations where riots and protests may occur. Here, we build on this work, specifically seeking to understand the types of features or land covers that these models are leveraging in their estimations.

Understanding the features identified by deep learning models is important for advancing research in the growing field of socioeconomic estimation from satellite data. Satellite imagery enables researchers to address topics that were previously difficult to measure, including estimating socioeconomic metrics such as wealth [6,7], predicting educational outcomes [8], human migration patterns [9], conflict-related fatalities [10], and economic devastation in conflict zones [11], predicting riots in urban areas [5], and conflict prediction [12]. Modern platforms like Planet [13] further enhance this capability by providing near-real-time global coverage, allowing for continuous monitoring and analysis.

However, a critical gap remains in our ability to interpret which specific factors within these models drive their predictions, limiting both transparency and trust in their outputs.

While significant attention has been devoted to explaining AI model outputs in the broader XAI community and even in the context of satellite-based models [14], there is a notable absence of literature specifically addressing the explainability of features driving socioeconomic outcomes derived from satellite imagery. Although numerous studies have demonstrated the capability of satellite imagery to predict factors such as household wealth or conflict [15,16], the underlying mechanisms remain largely unexplored. This knowledge gap is particularly significant given the increasing reliance on these predictive models for policy-making and resource allocation in developing regions [17]. The present study aimed to bridge this gap in the context of conflict studies by addressing two fundamental research questions, thereby contributing to the nascent field of XAI in satellite-based socioeconomic predictions:

RQ1. *How effective are class activation maps at localizing the regions of satellite imagery that are relevant for the classification of socioeconomic factors?*

RQ2. *Can land use be used as a proximate factor to understand the likely location of riot events in urban environments?*

To explore these questions, we employed a multi-faceted approach combining deep learning techniques with explainable AI (XAI) methods, focusing on the prediction of urban conflict from satellite imagery. Our methodology leveraged convolutional neural networks (CNNs) trained on high-resolution satellite imagery to predict the likelihood of riots and protests in urban areas. To explore RQ1, we applied Score-CAM (class activation mapping) techniques to generate heat maps highlighting the most influential regions within each image for the model's predictions and quantify the model degradation when these regions were removed. The goal of this approach was to first verify Score-CAM's ability to locate regions of satellite imagery relevant to classification. If Score-CAM was able to indicate regions successfully, the next step (RQ2) was to semantically define the features the network was identifying. For RQ2, we overlaid high-resolution land use categories onto the activation maps and then analyzed the correlation between specific land use types and areas of high activation. This analysis allowed us to determine whether certain land use categories were disproportionately influencing the model's predictions of urban conflict. This research aimed to evaluate how effectively current explainable AI (XAI) methods can predict events such as riots or protests, particularly in cases where satellite imagery does not contain clear, easily identifiable features that directly relate to these events.

This paper is structured as follows. Background information and the current research are discussed in Sections 1.1–1.4. In Section 2, the materials and methods are discussed, including dataset construction in Section 2.1 and a detailed methodology in Section 2.2. The results are presented in Section 3. Finally, Sections 3 and 4 present the discussion and conclusion.

1.1. Explainability in Machine Learning

One of the main critiques of neural networks and other machine learning techniques is their tendency to function as "black boxes", offering predictions without clear insight into the underlying features or relationships being exploited [18–22]. To address this opacity, researchers have developed various explainability methods that reveal which image features (such as objects or regions) are essential for a model's predictions. These methods include analyzing intermediate convolutional layers [23], inverting neural networks to reconstruct input features [24,25], and using class activation mapping (CAM) to highlight

important regions within an image [26]. While CAM methods are effective at identifying regions relevant to image classification, other approaches have been explored to enhance explainability. For example, attention mechanisms explicitly model the dependencies between features, as demonstrated by self-attention techniques [27]. Vision Transformers extend this idea by modeling attention across an image, though they are primarily designed for object-centric datasets [28]. However, the satellite imagery used in this research presents unique challenges: the features indicative of riots are not well-defined objects, and their resolution may vary across images. In this context, CAM-based visual heat maps offer a robust solution, capable of identifying relevant regions even under these uncertain and variable conditions.

Researchers' approaches to explainability, and the extent to which they aim to uncover the underlying behavior of models, vary significantly. Layer-wise Relevance Propagation [29] decomposes the prediction of a neural network to asses the relevance of input features at the pixel level, offering a detailed means of opening the "black box". In contrast, some researchers propose counterfactuals [30] that alter the inputs to neural networks to change the classifications, providing explanations without opening the "black box". In a review, Höhl et al. found CAM techniques to be one of the more frequently used explainability methods, particularly in land cover tasks [31].

Grad-CAM introduced the idea of examining training gradients to produce heat maps, visually identifying regions critical for classification [32]. Its success has inspired numerous extensions, such as Grad-CAM++, which improves the object localization accuracy [21], and SSGrad-CAM, designed to enhance the sensitivity to small features [33]. Other variations include Integrated Grad-CAM, which applies path integrals to gradients for refined analysis [34], and Score-CAM, which replaces gradients with activation-based weights to enhance stability [35]. A further refinement, Score-CAM++, uses logarithmic transformations to minimize noise from non-target elements in the activation map [36]. These advancements collectively aim to make neural networks more interpretable, bridging the gap between model performance and transparency.

For further information about the broader field of XAI, we recommend reading [31,37].

1.2. Explainability in Satellite Imagery Machine Learning Applications

The literature on applying explainability techniques to satellite imagery remains limited. The existing research predominantly focuses on well-defined identification tasks, such as detecting well-defined semantic classes like bridges, ports, stadiums, and airports [38] or counting objects such as airplanes in airport imagery [39]. According to Höhl et al., the three most common satellite imagery tasks where researchers apply explainability techniques are land cover mapping, agricultural monitoring, and natural hazard monitoring [31].

A variety of tools have been employed to address explainability in satellite data applications. Researchers have explored techniques such as Saliency Maps [40], LIME [41], and SHAP [42]. Saliency Maps have demonstrated effectiveness in ship detection tasks, outperforming other methods with a recall exceeding 93% [43]. SHAP has been used to visualize pixel-level correlations with classification outcomes in land cover tasks, achieving over 94% accuracy with a CNN [44].

Khan et al. compared Grad-CAM and Integrated Gradients in the context of satellite imagery, finding mixed performance depending on the target class [45]. Integrated Gradients, implemented via the Captum library [46], also offer a general framework for interpretability across a range of tasks beyond satellite imagery, such as text classification.

Explainability in satellite imagery poses several challenges that differ from those in traditional object-centric image datasets:

- Class overlap: The presence of multiple objects from the same class in an image, such as airplanes in an airport scene, complicates classification [47].
- Gradient noise: Models relying on gradient-based methods often struggle with the noisy gradients inherent in satellite data [35].
- Limited spectral information: Compared to traditional images, satellite imagery provides restricted pixel values and band ranges, limiting the variability of available information [48,49].

These domain-specific challenges underscore the complexities faced by researchers in achieving effective explainability for remote sensing data.

1.3. Lulc Used to Explain Human Behavior and Social Modeling in GIS

Another avenue to exploit satellite imagery to understand human behavior is to explore correlations between the land use and land cover (LULC) and social phenomena, such as conflict [50,51]. LULC analysis has established itself as a fundamental concept within the Geographic Information System (GIS) community, frequently employed to model human behavior and explain social phenomena [52,53]. This approach has been applied to various research domains, including the study of urban land expansion to understand and predict the global-scale impacts [54] and the integration of LULC data with population studies to model urban expansion [55], with ongoing efforts to establish connections between LULC patterns and socioeconomic processes [56]. In specific contexts, such as the use of pastoral lands in Ethiopia, the LULC and changes in land use have been utilized to explain conflicts between groups [50,51].

The application of LULC analysis in satellite imagery research has a rich history in the literature [57–61]. Various LULC studies have provided evidence to explain human activities and social interactions. For instance, the interaction between governmental policy and land use has been studied in the Mekong Delta [62], Columbia [63], and Inner Mongolia [64]. Furthermore, the LULC is frequently employed to model, explain, and understand urban growth and expansion across diverse global regions [65–68]. The LULC has also been used to understand the social dynamics of suburban regions [69,70].

Urban land use data, a specialized form of LULC information, have emerged as a powerful tool for understanding and modeling complex urban dynamics [71,72]. This approach has been applied to a diverse range of research areas, significantly enhancing our understanding of urban systems and human behavior within them. Traditional applications include studies on travel patterns, where researchers have explored the relationships between urban land use configurations and travel behaviors [73,74]. More recent investigations have revealed direct links between the urban land use mix and various urban phenomena, providing insights into how urban design influences human behavior and public health outcomes [75,76].

Recently, the incorporation of human-generated data, such as mobile phone records, with satellite imagery has enabled the highly accurate mapping of urban land use patterns [77]. Researchers have developed methods to model changes in the urban land use and associated human behavior with high degrees of accuracy [78]. These models not only provide a snapshot of the current urban conditions but also offer predictive capabilities for future urban development scenarios [79,80]. Such advancements in urban land use modeling contribute significantly to urban planning, policy-making, and sustainable development strategies [81].

While the links between the land use and human behavior have been explored in various contexts [52,53,56], there is relatively less focus placed on distinguishing between cases where relationships may be defined by specific objects—such as the presence of a car—rather than broader land cover categories. This distinction is crucial, as it may reveal more nuanced drivers of human behavior that are not captured by traditional land use classifications. In

urban environments, for instance, the presence of certain objects or structures might be more indicative of social dynamics than the general land cover type [75,77].

In the context of conflict studies, this differentiation becomes particularly relevant. While the LULC has been used to explain conflicts in some scenarios, such as disputes over pastoral lands in Ethiopia [50,51], it may not fully capture the complexities of urban conflicts. The urban landscape contains a multitude of features that could potentially influence conflict dynamics beyond what is typically captured in land cover classifications.

In this study, we aimed to explore the degree to which the land cover may provide an alternate, more semantically interpretable (as contrasted with deep learning) explanation of conflict in urban settings. By comparing the explanatory power of traditional LULC categories with the features identified by our deep learning model, we sought to understand whether conventional land use classifications are sufficient for predicting and explaining urban conflicts or if more granular, object-level features play a significant role. This approach not only contributes to the field of conflict studies but also challenges the conventional use of the LULC in explaining complex urban phenomena [54,55,62].

1.4. Understanding Conflict Through Satellite Imagery

Satellite imagery, combined with machine learning techniques, has become an important tool in conflict studies, offering new ways to analyze conflict dynamics, assess damage, and understand the socioeconomic impacts. Researchers have developed methods to leverage this technology for predicting protests and riots in urban environments [5]. Goodman et al. utilized convolutional neural networks (CNNs) and satellite data to predict conflict fatality risks in Nigeria, achieving an area under the ROC curve exceeding 75% [12]. Their earlier work also involved predicting non-permissive environments using moderate-resolution imagery [10], laying the foundation for broader applications of satellite imagery in identifying unstable regions.

In post-conflict damage assessment, deep learning models have been effective in detecting building damage. Xu et al. and Mueller et al. demonstrated the utility of these methods across various conflict zones [82,83], while Nabiee et al. reported accuracies of between 75% and 92% in identifying damaged buildings during the Syrian civil war [84].

Beyond immediate conflict zones, Murillo-Sandoval et al. employed satellite imagery to monitor land cover changes related to illicit activities in Colombia, successfully identifying 96% of the overall changes, including the expansion of illegal coca farming [63]. Eklund et al. used similar techniques to study agricultural land use changes in territories controlled by the Islamic State in Iraq and Syria, achieving 80% accuracy in tracking these changes [85].

In this piece, we extend this literature to riots and protests in urban areas, with a focus on explaining the underlying factors which drive such events.

2. Materials and Methods

2.1. Data and Labeling

We developed a dataset according to the process described in [5], combining data from multiple open-source repositories with commercially available satellite imagery. The primary motivation of the dataset construction was to download satellite images predicated on known riot and protest events. These satellite images were used to generate clipped riot images and clipped non-riot images. For a more in-depth description of the process, please see [5].

2.1.1. Data

Satellite imagery data were the primary source of information used to train the neural network and classify locations as either an urban area which experienced a riot or a “non-

riot" location at which no such event occurred. *Planetscope* images were downloaded from Planet [13] that corresponded to 24–48 h prior to a known riot event. *Planetscope* provides 3–4 m resolution images. A more detailed description of the imagery products available is shown in Table 1. The *Planetscope* constellation has changed over time, including the type and number of bands available for download. Only the visual bands (RGB) of locations that had 50% or less cloud cover were retrieved.

Table 1. This table presents the technical details of the satellite imagery, based on the *Planetscope* sensor [86]. The details for each image correspond to the generation of satellite used to image the location.

Instrument	Image Area	Availability	Wavelength (nm)		
			Red	Green	Blue
Dove Classic	25 × 11.5 sq km	July 2014–April 2022	590–670	500–590	455–515
Dove-R	25 × 23 sq km	March 2019–April 2022	650–682	547–585	464–517
SuperDove	32.5 × 19.6 sq km	March 2020–present	650–680	547–583	465–515

The Armed Conflict Location Event Data Project (ACLED) [87] database provided the location of known riots. The ACLED is an open-source database containing information about conflict from around the globe. While the ACLED database contains over 1.5 million entries, we only considered entries that were from riots and/or protests (referred to simply as riots in this paper). Additionally, only riots that were georeferenced to a specific neighborhood within an urban city were included. This level of geographic precision provided better localization during the clipping of satellite images during data labeling.

An additional source of data was the Degree of Urbanisation (DEGURB) dataset developed by the European Commission's Joint Research Center [88]. These data classify the entire surface of the globe in terms of the population density and are released every 5 years. For our purposes, we only considered areas with a population density of over 300 inhabitants per square kilometer, seeking to only include urban locales [89]. The DEGURB data created a binary mask for the globe, allowing us to subsequently restrict satellite imagery information to only include regions with 300 or more inhabitants per square kilometer. An example of this binary mask is shown in Figure 1.

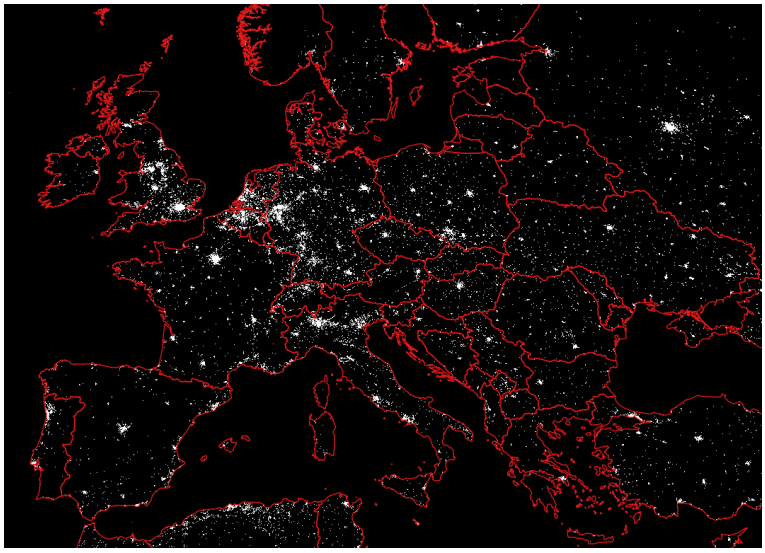

Figure 1. A portion of the DEGURB data, highlighting the areas of the world that were considered urban in our dataset. The DEGURB defines urban regions as those with a density of more than 300 inhabitants per km [89]. Red lines represent country-level boundaries [90].

The final data source was the urban land use (ULU) data introduced by Guzder et al. [91]. These data provided a 7-class urban land use map derived using Sentinel-2 satellite imagery. Guzder et al. constructed this dataset with a 5 m spatial resolution for approximately 4000 cities across the globe. The seven semantic classes they used were open space, non-residential areas, atomistic settlements, informal land subdivisions, formal land subdivisions, housing projects, and roads [91]. These seven classes allow for easy delineation between open spaces and built-up areas higher in the hierarchy and more detailed analysis lower in the hierarchy when labeling the four different types of residential areas. An example of this land cover classification is shown in Figure 2.

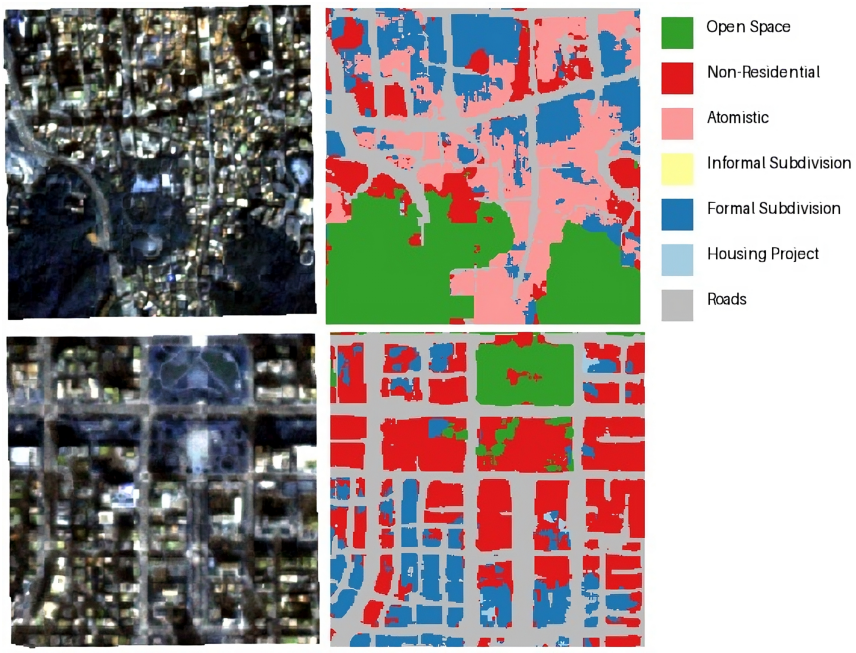

Figure 2. These are examples of the ULU data from Guzder et al. [91]. The satellite images on the left of both rows are from two different neighborhoods in South Korea. The land use maps in the center of each row show the ULU data associated with each neighborhood. The legend listing the 7 classes as described by Guzder et al. is on the right. Imagery ©Planet Labs PBC 2023. All rights reserved.

Guzder et al. reported that, at a 5m spatial resolution, their 6-class (omitting roads) product had an overall accuracy of 71%. They reported 85% accuracy in identifying roads, present in the 7-class dataset [91]. This data source served as a potential explanatory variable for what features the neural network learned to classify as riot and non-riot images. For example, these may have included factors such as the amount of the area that was covered by roadways or if there was a connection between the types of residential areas present that correlated with riots in a neighborhood.

2.1.2. Labeling

To construct a labeled dataset for use in training a deep learning model, we integrated multiple data sources. Riot events were aggregated using the information from the ACLED [92], specifically focusing on the location and date of each event. Satellite imagery was then retrieved from Planet [13], corresponding to a 24–48 h window prior to each riot. This time frame was chosen to capture the environmental and infrastructure features present before the riot, rather than attempting to capture images of crowds or the riot itself. The goal was to identify potential precursors or patterns that may precede a riot.

For each event, the neighborhood where the riot occurred was isolated, and a 1-square-kilometer image was clipped from the corresponding satellite tile. This image was labeled as a "riot" instance in the dataset. To generate control cases, we used the remaining portion

of the satellite tile outside the riot-affected area. Using DEGURB data [88], we ensured that all control cases were drawn exclusively from urban regions within the tile. From these urban areas, a single 1-square-kilometer clipping was randomly selected and labeled as a "null" (non-riot) instance in the dataset. This labeling methodology allowed us to generate a balanced dataset that had riots and non-riots equivalently represented in it. Additionally, any imagery, sensor, or seasonal biases were equally represented across classes.

In the final dataset, there were 16,274 satellite images from 24 countries (representing countries with a minimum of 500 or more images). Using a test and train split of 80% for training and 20% for testing for each country, we constructed a training set where each country was represented at least 400 times and a validation testing set where each country was represented at least 100 times. Details are provided in Table 2.

Table 2. There were 32,548 clipped images in the dataset. Half of these were from riots, and half were null clippings. Only countries that had at least 500 images were included. A total of 20% of each country's images was withheld from training and testing and used in validation.

Country	Images	Country	Images	Country	Images
South Korea	7494	Pakistan	2622	Iran	2334
Lebanon	1656	Israel/Palestine	1572	China	1550
South Africa	1480	Chile	1302	Japan	1256
India	1148	Brazil	1112	Bangladesh	1092
Ukraine	924	Thailand	890	Italy	728
Russia	668	Indonesia	678	Venezuela	648
Greece	634	Yemen	604	Taiwan	562
U.K.	566	Peru	522	Iraq	506

The wide range of countries represented in the dataset allowed us to explore the capability of explanatory AI techniques at a global scale and further enabled tests as to the spatial variation in the explanatory capabilities.

2.2. Methods

Our methodological approach followed a multiple-step process:

1. A convolutional neural network (ResNet18) was fit to the 32,548 images labeled as the location of a "riot" or "no riot", following the procedures outlined in [5]. We recorded the global accuracy and country-level accuracy for the model.
2. Class activation mapping (Score-CAM) was applied to every image, generating a heat map localizing the regions of each image important for its classification.
3. A threshold value was set to filter areas identified by Score-CAM as being critical for the classification; these areas were removed from each image (all values were set to 0), and then each image was classified as the location of a riot or non-riot a second time. We contrasted the revised accuracy (after critical areas were removed) with the original accuracy to determine the efficacy of Score-CAM in localizing where important features existed within the image.
4. Using the same localized regions identified by Score-CAM, we tested if there were meaningful land cover distinctions between those regions and the remainder of each urban tile, testing if the land cover could provide a useful proximate feature for identifying riot locations.

Each of these steps is described in more detail below and is summarized in Figure 3.

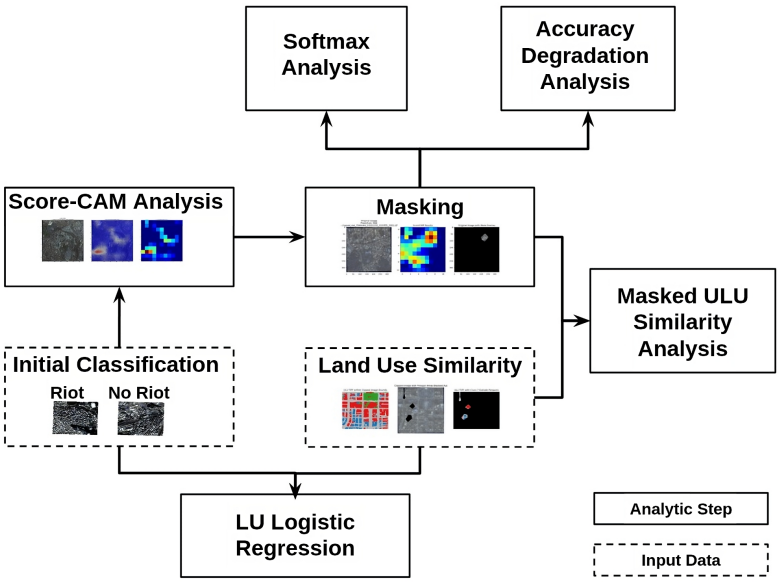

Figure 3. A synopsis of the analysis conducted after training ResNet18. The data generation and training were similar to those in [5]. After training, a Score-CAM analysis determined the regions of interest to mask. The masking relevance was evaluated through accuracy and softmax analysis. Finally, the land use was explored as an explanatory variable to understand the deep learning classification rationale. The figure was adapted from [5]. Imagery ©Planet Labs PBC 2023. All rights reserved.

2.2.1. Neural Network Training

As in [5], ResNet18 was trained with 32,548 images (16,274 from riots and 16,274 from non-riots) and validated with a testing dataset consisting of 6490 images (3245 from riots and 3245 from non-riots). These images were from 24 countries, with each country having at least 500 images (see Table 2). By selecting countries from around the world, with a minimum of 500 images each, we sought to mitigate biases from any one country or region impacting the results. Prior to training on this dataset, ResNet18 was pretrained with weights from ImageNet. After training on 80% of the data, the performance of the network was tested on the remaining 20% for validation. The network achieved over 89% accuracy, demonstrating the ability to distinguish between neighborhoods where riots will occur and neighborhoods where riots will not occur. The trained ResNet weights were saved and constituted the trained neural network used throughout this work.

2.2.2. Score-CAM-Informed Masking

The explicit research question we sought to answer was the following: *How effective are class activation maps at localizing the regions of satellite imagery that are relevant for the classification of socioeconomic factors?* Here, we specifically tested Score-CAM, a technique which uses the weights of a selected convolutional layer to generate a heat map visualizing the portions of the image relevant to classification. This research followed the implementation presented in [35], using the weights of the last convolutional layer to generate the Score-CAM results, an example of which is displayed in Figure 4.

Integral to exploring explainability was determining how effective Score-CAM was at localizing the features which were important in driving the classification accuracy across each satellite image. By understanding the performance of Score-CAM in this application, we could gain insight into the broader explainable AI field. Different thresholds for Score-CAM were evaluated to focus efforts on the regions of the image that mattered the most in classification. The Score-CAM weights from the final convolutional layer were used, which ranged from 0.0 to 1.0. The higher the threshold value selected, the smaller the regions

of interest. A threshold of 0.0 would include most of the image. A Score-CAM threshold just under 1.0 would exclude almost none of the image. In this paper, the Score-CAM threshold was incremented by 0.05 from 0.0 to 1.0, each time masking values above the given Score-CAM threshold. This process was repeated to determine the optimal threshold for analysis, prioritizing a value that identified a sufficiently large area for semantic definitions while ensuring that masking out the identified important regions resulted in the model's prediction shifting from a riot classification to a non-riot case. This shift confirmed the successful localization of critical pixels influencing the prediction. An example of this process can be seen in Figure 5.

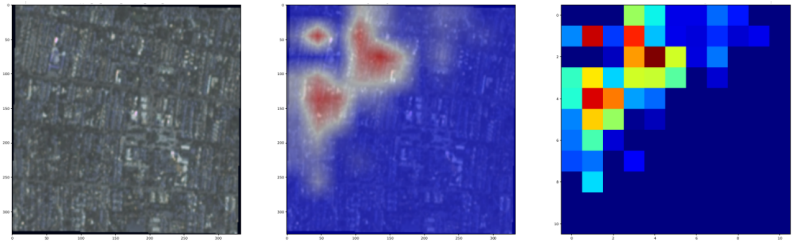

Figure 4. An example of a riot image on the left, with Score-CAM overlaid on top of the riot image in the center and the Score-CAM results alone on the right. Imagery ©Planet Labs PBC 2023. All rights reserved.

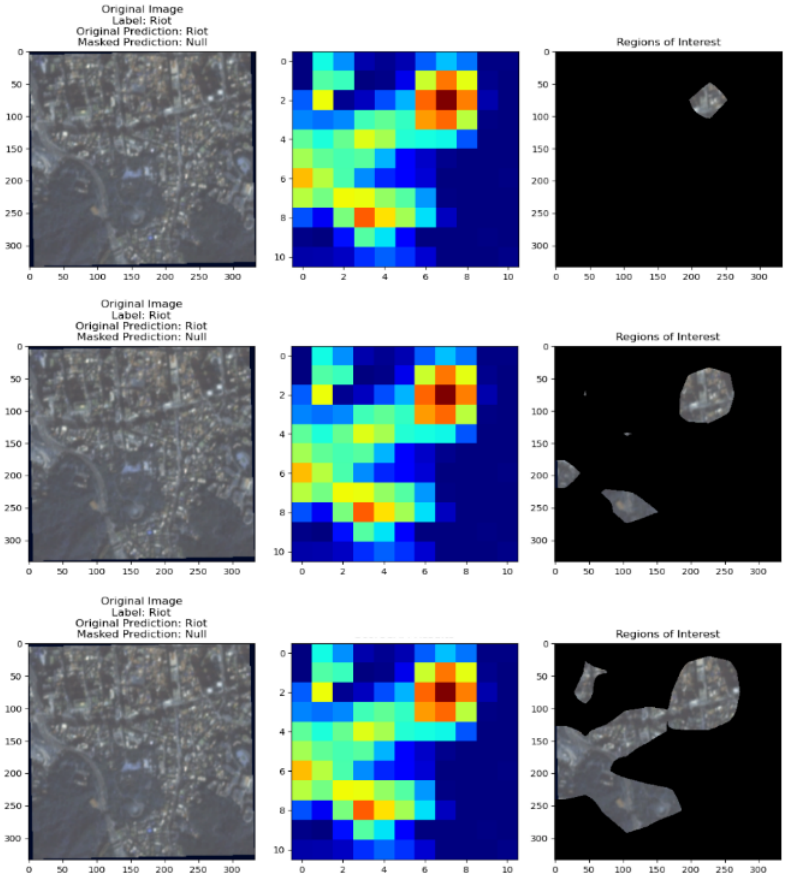

Figure 5. An example of finding regions of interest across different Score-CAM thresholds. Regions above the threshold are visible, while regions below the threshold are masked. During implementation, the regions of interest were obscured from ResNet during evaluation. The top row is an example of a threshold of 0.8, the middle row is an example of a threshold of 0.6, and the bottom row is an example of a threshold of 0.4. The satellite images on the left of each row are identical, as well as the Score-CAM results. The regions of interest shrink as the threshold increases. Imagery ©Planet Labs PBC 2023. All rights reserved.

2.2.3. Evaluating Score-CAM's Effectiveness

After masking regions above each Score-CAM threshold (ranging from 0.0 to 0.95), ResNet's performance was evaluated by measuring the accuracy, precision, recall, true positive rate, and false positive rate. As masking reduces the available information, a decline in the classification performance was expected as larger quantities of each image were masked. We compared the original and masked metrics using a forward pass through the same network to assess the significance of the masked features for accurate predictions.

For each observation, the ResNet model generated a softmax output that reflected the likelihood of a given location containing a riot. Although these outputs do not represent true class probabilities, they can be interpreted as confidence scores, with values closer to 1 indicating higher confidence. In the binary classification scenario (riot vs. non-riot), values near 0.5 suggested low confidence. While softmax outputs are known to exhibit a bias toward higher values [93,94], they remain useful as comparative metrics for analyzing changes in the classification confidence. In this work, we utilized softmax outputs to assess the model's confidence in a given classification by processing the same image twice: once with all the data and a second time with masked information.

In this controlled comparison, where no other factors changed, the direction of the shift in softmax values represented the change in the model's confidence regarding the class inclusion. Softmax outputs have been used in a range of applications to highlight shifts in confidence and explain misclassifications. For instance, Moon et al. [95] demonstrated the use of confidence-aware learning by leveraging softmax probabilities to detect shifts in the classification confidence, emphasizing the importance of probabilistic outputs for model interpretability. Hendrycks and Gimpel [96] similarly showed that correctly classified examples tend to exhibit higher maximum softmax probabilities than misclassified ones, providing a means of detecting errors. Rozsa et al. [97] further explored the adversarial robustness of deep neural networks by comparing softmax layers with Openmax, illustrating how softmax outputs can be used to assess the model's confidence and vulnerability to adversarial perturbations. In the context of adversarial attacks on neural networks, including ResNet101, Sen et al. [98] utilized the softmax results post-attack to demonstrate how neural networks lose confidence and misclassify images.

To determine changes in confidence, the original image softmax was compared to the masked image softmax results. Specifically, we subtracted the softmax estimated with masked imagery from the original softmax value, aiming to quantify the shift in the model's confidence after masking the regions of interest. For example, if an image classification had a softmax result of [0.99, 0.01], where the first value in the output represents a non-riot and the second value represents a riot, the network would classify the image as a non-riot. If after masking, the softmax result was [0.71, 0.29], the image would still be classified as a non-riot, but there would be less confidence in the classification. In this example ($0.99 - 0.71 = 0.28$), we would project a decrease in confidence (noting that the absolute decrease is skewed due to the propensity for softmax values to approach maximums, we interpret this as evidence of the directionality of a shift in the model confidence, rather than focusing on the absolute shift). In this analysis, positive values indicated decreasing confidence, while negative values represented increasing confidence. This was used as a secondary metric to assess the ability of Score-CAM to isolate important features across different thresholds.

2.2.4. Land Use Analysis

In this analysis, we engaged with our second research question: *Can the land use be used as a proximate factor to understand the likely location of riot events in urban environments?* If the land cover serves as a proximate factor, then deep learning may be unnecessary to localize

intra-urban riot events. Conversely, if the land cover is not correlated with important features which predict riots, then deep learning may reveal important patterns that could shed light on the relationship between the urban form and conflict.

The ULU data (described in Section 2.1.1) were used, for each image, to calculate two pixel counts. The first was, for each of the 7 classes, the number of pixels which fell within the original image. The second was all the pixels which fell inside the regions of interest identified by Score-CAM. By contrasting these two values, we sought to establish the potential of the land cover for the localization of riot events.

Because the original image and the regions of interest were not equivalent in size (recognizing that the geographic area of thresholded regions may vary across every image), each vector was normalized as shown in Equations (1) and (2). Specifically, we divided the count for each class, i, by the total count of pixels in the vectors:

$$\hat{O}_i = \frac{O_i}{\sum_{i=0}^{6} O_i} \tag{1}$$

$$\hat{R}_i = \frac{R_i}{\sum_{i=0}^{6} R_i} \tag{2}$$

where $\hat{O}$ represents the normalized vector of the original image and $\hat{R}$ represents the normalized vector of the regions of interest.

We implemented two tests to establish the degree to which the land cover may be a useful explanatory variable for identifying intra-urban riots. First, we implemented a linear regression in which the proportion of each class present in the original image, $\hat{O}$, was used as a variable to predict whether the image was from a riot or a non-riot. Second, we tested if the land use could distinguish between the features CAM-based techniques identified as important and the rest of the image. If the regions identified by Score-CAM were significantly different from the full image in terms of their land use proportions, land use data might explain the features deep learning techniques had learned.

To determine the similarity or dissimilarity between the images and Score-CAM-identified regions, we considered the L1 similarity, L2 similarity, and Cosine similarity. These three different similarity measures evaluated how similar or dissimilar the original images were to the regions of interest; i.e., a perfect similarity score would indicate that there was no difference between the Score-CAM-identified region and the broader image, suggesting that the land cover was insufficient to describe the features being detected by the deep learning algorithm. Given the 7 ULU classes, the L1 similarity for two perfectly aligned vectors would have a value of 0, and two orthogonal vectors would have a value of 7. The L2 evaluation would have a value of 0 for two perfectly aligned vectors, while two orthogonal vectors would have a value of $\sqrt{7} \approx 2.64$. In a Cosine similarity evaluation, since the vectors are normalized, two perfectly aligned vectors would have a value of 1, while two orthogonal vectors would have a value of 0.

3. Results

Our first goal was to test the capability of Score-CAM to identify key features of importance in satellite imagery. Prior to masking, our trained ResNet classified images with an accuracy of over 89%. After masking an average of 1.6% of each image using a Score-CAM threshold of 0.8 (see Section 3.1 below for more details on why this threshold was selected), the accuracy dropped to 68% (see Table 3). This illustrates that the features which were being masked by Score-CAM were representative of those that were critical to modeling predictions, as removing them resulted in a significant degradation of the

model performance. In addition to a drop in accuracy, there was a decrease in the model's confidence in image classification, indicated by a shift in the softmax scores (see Figure 7).

As a second objective, we sought to establish the degree to which traditional land cover metrics can be used to identify important features which distinguish riot or non-riot locations. We tested this using two approaches: a linear model, and by contrasting the similarity of the ULU inside and outside of regions identified as important by Score-CAM. Both results suggested that the ULU alone is insufficient to identify protests in urban environments; these findings are examined in the following subsections.

Table 3. This table compares the accuracy and other relevant metrics between the original image and the image after masking the regions of interest. After masking the regions of interest in the images, the neural network's ability to perform image classification was significantly degraded. These original image results match the findings in [5].

	Original Image	**Masked Image**
Accuracy	89.41%	68.54%
Precision	89.38%	79.91%
Recall	89.46%	49.52%
F1	89.42%	61.15%
True Positives	2903	1607
False Positives	345	404
True Negatives	2900	2841
False Negatives	342	1638

3.1. Threshold Analysis

To determine the optimal Score-CAM threshold for identifying critical features in satellite imagery, we conducted a threshold analysis in which we iteratively tested different thresholds in increments of 0.05. Figure 6 shows the results of this analysis. In this figure, the red line represents the false negative rate (the frequency of incorrectly classifying a riot as a non-riot), while the green line represents the true positive rate (the frequency of correctly classifying a riot). The intersection of these lines at 50% represents the threshold (in this case, 0.8) where the model's ability to distinguish riots from non-riots is significantly impacted. This threshold serves as an upper limit to where Score-CAM continues to provide effective localization. For instance, Figure 6 demonstrates that selecting a threshold higher than 0.8 results in Score-CAM identifying very few pixels, which leads to a minimal degradation of the model's performance, indicating that the selected pixels are not critical to the model's predictions. At the 0.8 threshold, the model's performance deteriorates to the extent that false positives outnumber true positives. This suggests that (a) the pixels identified by Score-CAM are indeed sufficient to degrade the model's performance when masked and (b) the selected image regions genuinely represent the most critical pixels driving the classification.

To evaluate how effective a threshold of 0.8 was in isolating riot-relevant information from images, we compared the original accuracy of the model to the accuracy after applying the mask. By using a threshold of 0.8, we masked areas of the image where the Score-CAM values were 0.8 or higher, which corresponded to the most important features identified by the model. This comparison allowed us to assess how much information relevant to riot classification was captured by these high-activation regions. The results of this analysis are shown in Table 3, illustrating the impact of masking on the model performance.

The results indicate that applying a 0.8 threshold to mask the regions of interest led to a substantial reduction in performance, decreasing the accuracy from nearly 90% to approximately 68%, while only masking 1.6% of the image. At this threshold, the true positive and false negative rates matched, indicating that the neural network was as likely to correctly

classify a riot as it was to not classify a riot. This outcome suggests that the regions identified by Score-CAM contained features important for accurate riot prediction, but not in a consistent manner, i.e., that Score-CAM could provide a useful explanatory tool for some aspects of the estimation of riot locations from satellite imagery but failed in other areas. These results for each country are discussed later, with the accuracies displayed in Table 5.

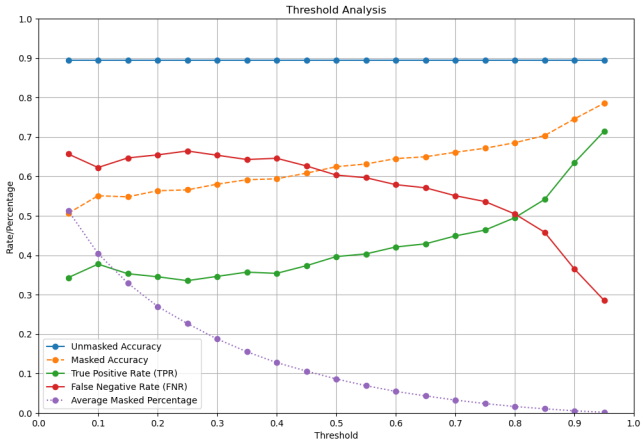

Figure 6. The results of the threshold analysis, with the threshold along the *x*-axis and the rate/score along the *y*-axis. The unmasked accuracy was 89%, displayed in blue. The accuracy after masking increased as the threshold increased, shown in orange. The true positive rate increased as the threshold increased, shown in green. The false negative rate decreased as the threshold increased, shown in red. The average masked percentage in purple is the amount of the image on average that was masked at the given threshold.

3.2. Softmax Analysis

Across the full dataset, after masking the regions of interest, the network was less confident in classifying images, as shown in Figure 7. Across all 6490 images, we observed a mean shift of 0.25, indicating less confidence in the classification after masking. There was a bimodal distribution for the shift in confidence across all images. While roughly 3000 images, nearly half of the data, showed no change in confidence, there is a significant concentration near 0.95 in Figure 7. Further exploration of the bimodal nature of the softmax shift is displayed in Figure 8. By analyzing the behavior of the true positives and the true negatives specifically, we were able to gain insight into the impact of masking images that ResNet was correctly classifying.

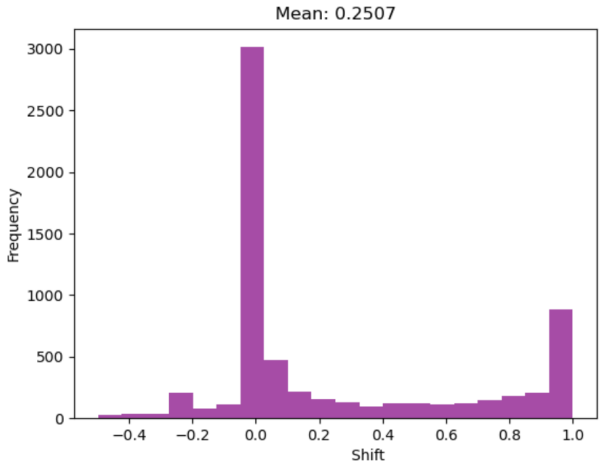

Figure 7. The softmax shift after masking. The softmax value associated with the original prediction minus the same value after masking created a softmax shift for every image. Positive values indicate a less confident classification after masking.

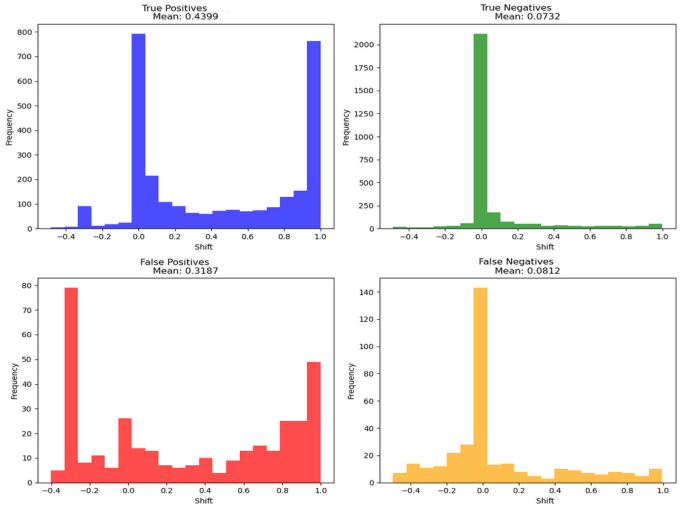

Figure 8. The softmax shift after masking, exploring the different cases in our confusion matrix. True positives (correctly predicting a riot) are in the upper left, shown in blue. True negatives (correctly predicting a non-riot) are in the upper right, shown in green. False negatives (predicting a riot in a non-riot image) are in the bottom left, shown in red. False negatives (predicting a non-riot in a riot image) are in the bottom right, shown in gold. Note that due to the different population sizes of the four cases, the y axis is not equivalent across the four charts.

The bimodal nature of the distribution in Figure 7 was driven by distinct shifts in the classification confidence for true positives and true negatives. After masking, approximately half of the true positives (upper left panel in Figure 8) showed a substantial reduction in confidence, while the other half exhibited no change. In other words, after masking, approximately half of the correctly classified riots were now classified as non-riots. This pattern aligns with the results in Figure 6, where the true positive and false negative rates intersect, suggesting that Score-CAM effectively identified key features in some cases but failed to do so consistently.

In contrast, true negatives (upper right panel in Figure 8) maintained a high classification confidence after masking, indicating that the regions identified by Score-CAM had minimal influence on the model's ability to classify non-riot images. Simply stated, masking correctly classified non-riot images did not change the model's behavior. This asymmetry underscores the mixed performance of Score-CAM in localizing relevant features in satellite imagery, particularly for more complex scenarios like identifying riot-specific patterns.

3.3. Ulu Data as an Explanatory Variable

Here, we tested if using ULU data to predict the location of a protest would be a suitable replacement for deep learning techniques. Any image that had 100% missing data or open space was excluded from the ULU analysis. Of the 6490 images in the dataset, 949 images did not overlap with the ULU dataset, leaving a dataset of 5541 images.

A linear regression was fit using these observations, using the percentage of each class present in the vector of the unmasked image, $\hat{O}$, as outlined in Section 2.2.4. This resulted in an r^2 value of 0.1017. This can be interpreted as the class percentages of the pixels explaining only 10% of the variance in the prediction of riot images. This result could have been driven by many factors, including nonlinearity in the relationship, a lack of correlations between the underlying drivers of conflict and the land cover, or other factors.

Using a Score-CAM threshold of 0.8 (as identified in Section 3.1), we isolated key features of importance within each image to compare urban land use (ULU) data from the original image to those of the masked regions associated with the image classification. This process is displayed in Figure 9. We then sought to establish if the areas inside of each

threshold-selected region were unique relative to the broader image in terms of the ULU. The similarity was evaluated using the three different similarity metrics discussed earlier.

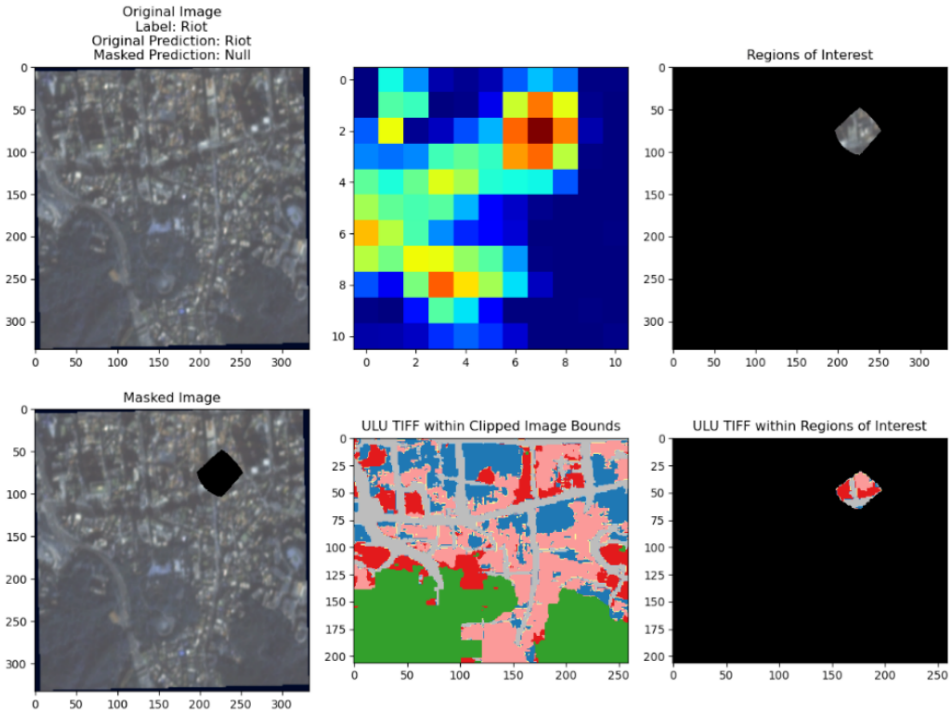

Figure 9. Satellite images were classified based on the features visible in the top left frame. Score-CAM indicated the regions of interest as displayed in the top center frame. The regions of interest were isolated based on the Score-CAM threshold; this is displayed in the top right frame. The masked image was evaluated, obscuring the regions of interest as shown in the bottom left. The ULU data for the full satellite image, shown in the bottom center, were compared to the ULU data for the regions of interest, shown in the bottom right. Imagery ©Planet Labs PBC 2023. All rights reserved.

The histograms of the values for each similarity score are displayed in Figure 10. The summary statistics for the similarity scores are shown in Table 4.

For each of the three similarity measures, a one-sided t-test was conducted to determine whether the distributions of the similarity scores indicated significant differences between the full images and the Score-CAM-identified locations of the features of importance. In all cases, the tests showed no significant difference in the means, suggesting that the locations of the features of importance were generally similar—in terms of the land cover—to the full images.

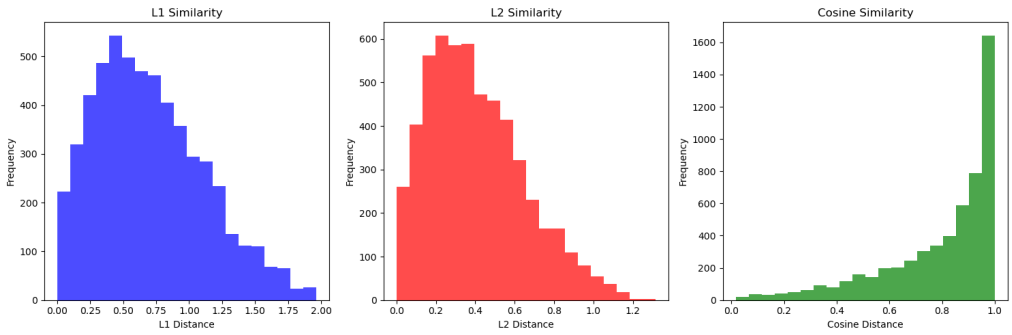

Figure 10. The frequencies of the L1 similarities are shown in blue on the left. The frequencies of the L2 similarities are shown in red in the center. The frequencies of the Cosine similarities are shown in green on the right.

Table 4. The results of different similarity scores. All three similarity measures indicated that the regions highlighted by Score-CAM were not significantly different from the full image.

	L1 Similarity	L2 Similarity	Cosine Similarity
Mean	0.7004	0.4046	0.7946
Median	0.6473	0.3682	0.8753
Standard Deviation	0.4143	0.2447	0.2157
Null Hypothesis	$\mu_0 \geq 3.5$	$\mu_0 \geq 1.3$	$\mu_0 \leq 0.5$
T Statistic	−502.99	−272.37	101.66
p Value	1	1	1
Result	Similar	Similar	Similar

3.4. County-Level Analysis

The dataset was constructed to allow for the country-specific analysis of the masking accuracy, softmax shifts, and ULU similarities; all contrasts are shown in Appendix A.1. Of particular note is that, when masking portions of the image, the model degradation was not uniform across all 24 countries in the dataset. The country-specific decreases in accuracy are shown in Table 5.

Table 5. Comparison of unmasked accuracy, masked accuracy, and change in accuracy by country.

Country	Unmasked Accuracy	Masked Accuracy	Change in Accuracy
United Kingdom	0.8393	0.5357	−0.3036
Taiwan	0.8214	0.5179	−0.3036
Ukraine	0.9185	0.6467	−0.2717
Israel/Palestine	0.8949	0.6338	−0.2611
Indonesia	0.8657	0.6194	−0.2463
South Africa	0.8243	0.5845	−0.2399
Bangladesh	0.8899	0.6514	−0.2385
Venezuela	0.8906	0.6563	−0.2344
South Korea	0.9212	0.6903	−0.2310
India	0.8860	0.6667	−0.2193
Thailand	0.8708	0.6517	−0.2191
Brazil	0.8514	0.6351	−0.2162
Pakistan	0.9256	0.7099	−0.2156
Italy	0.8819	0.6806	−0.2014
Yemen	0.7833	0.5833	−0.2000
Chile	0.9115	0.7231	−0.1885
Lebanon	0.9455	0.7576	−0.1879
Peru	0.7788	0.6058	−0.1731
China	0.9000	0.7290	−0.1710
Iraq	0.9100	0.7500	−0.1600
Japan	0.8560	0.7200	−0.1360
Greece	0.8730	0.7460	−0.1270
Iran	**0.9442**	0.8219	−0.1223
Russia	**0.7803**	0.6667	−0.1136

This table is ordered by each country's change in accuracy. As an example, masking caused the United Kingdom's accuracy to decrease from 83.8% to 53.6%, the largest shift in the dataset. The relatively high unmasked accuracy indicates that the neural network was particularly effective at learning what features predicted riots in the United Kingdom. This contrasts with Iran and Russia, which had the two smallest changes in accuracy, indicating that Score-CAM's output struggled more to distinguish riot features.

4. Discussion

4.1. Feature Detection in the Context of Socioeconomic Outcomes

This study's novelty is in its investigation of a socioeconomic outcome—riot activity—without predefined feature correlates; it is not clear which specific features in satellite imagery should or should not align with riot occurrences across different geographic regions. This ambiguity creates a unique opportunity to uncover new insights into underlying socioeconomic processes through satellite imagery, contingent on the ability to interpret the features identified by deep learning models.

This study demonstrates that deep learning techniques are well suited to identify specific imagery features that are associated with riot predictions. When as little as 1.6% of the image was masked from classification, there was a significant degradation in performance. Further, the directionality of shifts in our true and false positives outlined in Section 3.2 indicates that the neural network had learned to detect features associated with riots specifically, as opposed to distinguishing between riots and non-riots. The performance shift observed in true positives that is absent in true negatives in Figure 8 supports this claim. There was no observed degradation in the classification of non-riots, indicating that the network was initializing images as showing a non-riot and the detection of riot features changed the classification to a riot.

Deep learning appears to be particularly well suited for this classification task, especially in contrast to traditional GIS-based methods such as land use analysis. A land use pixel-based linear regression was only capable of accounting for 10% of the variance in predictions. Furthermore, when the relevant regions were isolated through deep learning, the traditional GIS land use analysis techniques failed to detect any significant differences in the land cover proportions between relevant regions and the broader images. Convolutional neural networks and class activation mapping were able to detect features that, when masked, degraded the performance from 89% to 68%, but land use analysis considered the masked regions as similar to the surrounding terrain.

While the direct causes of riots are not clear at this time, deep learning demonstrates the potential for the discovery and definition of the features used in successful classification. Traditional GIS techniques, such as the land use, do not provide the same capabilities that deep learning has demonstrated in the prediction of conflict locations. This underscores the fact that the complexity and heterogeneity of satellite imagery—where objects of interest can be obscured by atmospheric conditions, the topography, or seasonal changes—make universal model performance more difficult to achieve. This diversity emphasizes the need for robust domain adaptation strategies or additional fine-tuning when transferring a model to different locations or times of year. Score-CAM can help identify whether the model focuses on relevant features (e.g., certain spectral signatures of vegetation) or is distracted by domain-specific artifacts (e.g., shadows, clouds, or noise), thereby clarifying how well the model generalizes.

4.2. Limitations of Score-CAM and CAM Approaches

Score-CAM is capable of identifying some of the regions of interest within a satellite image, but it does not provide any semantically meaningful results. Without semantic definitions, it is difficult to understand what the features are or why they are relevant to classification. For example, while it may be notable that a given type of building is correlated with riots, without knowing the specific use of the building, it can be difficult to draw conclusions as to why.

A second challenge is that differences in the results when using Score-CAM in satellite imagery applications can arise from several factors. First, satellite images vary widely in the resolution, spectral bands, and illumination conditions, which can lead to domain shifts

that challenge a model's ability to generalize. A model trained on one sensor type, for example, may not transfer well to data gathered from another sensor with different spectral characteristics or spatial resolution. Such disparities can manifest as inconsistent or less meaningful Score-CAM visualizations, since the model's learned feature representations may not align with the new domain.

Score-CAM and other CAM techniques primarily focus on the training and evaluation of object-centric image datasets, such as ImageNet [35]. This application is fundamentally different to using satellite imagery. When predicting protests, satellite imagery is likely to contain many relevant objects in a single image that can differ greatly in size. As an example, a satellite image might contain a road that runs the length of the image, with multiple vehicles on the road which are much smaller in size. Furthermore, the context of vehicles on a road might be important to classification, but Score-CAM's ability to understand this context is not well understood at this time. It is unknown how these differences in the image type and object scale are handled with CAM techniques.

The multi-object nature of satellite imagery further complicates this analysis. The dataset comprised images of cities, each containing multiple elements such as roads, buildings, and parks. Figure 11 provides an illustrative example of this complexity. In this image, multiple roads and buildings are present, including several intersections. Interestingly, Score-CAM identifies only one road intersection as relevant to image classification, while disregarding others within the same image. This selective highlighting of features by Score-CAM raises intriguing questions about the algorithm's decision-making process in complex, multi-object scenarios. It suggests that certain instances of a feature type may be more influential in classification than others, even when they appear visually similar. This observation underscores the nuanced nature of feature relevance in satellite imagery classification and highlights the potential for further investigation into the criteria by which Score-CAM determines feature importance in such complex visual environments. This may include contextual factors which are currently not captured by available XAI approaches.

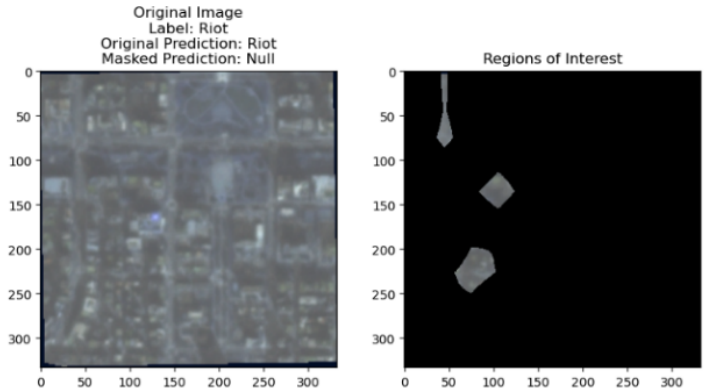

Figure 11. An image from a riot on the left and a masked image on the right. The regions below a Score-CAM threshold of 0.8 are masked. Masking this image resulted in the classification shifting from a riot to a non-riot. Imagery ©Planet Labs PBC 2023. All rights reserved.

Score-CAM demonstrates limitations in consistently identifying the features the deep learning model has successfully learned. As shown in Figure 8, approximately half of the true positives lost confidence after masking, indicating that relevant riot features were successfully localized to just 1.6% of the image. In contrast, the other half showed no change in confidence, suggesting that key features were not identified by Score-CAM. This inconsistency may have stemmed from features that exceeded the masked area or relied on contextual or spatial relationships that Score-CAM cannot capture. While the exact reasons

for these challenges remain unclear, they highlight the need for further research to improve explainable AI methods for satellite imagery.

The limitations of CAMs in explaining satellite data are most evident in the country-specific results of masking, as shown in Table 5. The inconsistent impact of masking on the accuracy highlights this issue. For example, Score-CAM effectively identified relevant features in the United Kingdom, resulting in the largest accuracy drop (from 83.9% to 53.6%). However, in Iran, where ResNet achieved a higher initial accuracy (94.4%), Score-CAM had a minimal impact, with the accuracy only decreasing to 82.2%.

As displayed in Table 5, the country-level results in Section 3.4 were not consistent for every nation in the dataset. In particular, ResNet struggled to accurately classify images from Russia. One example of this is shown in Figure 12, where the network correctly classified an image from a riot in Moscow prior to and after masking. This indicates that Score-CAM was not able to identify the features that drove classification; i.e., if the correct features were selected by Score-CAM, we would expect the classification to flip to "non-riot" after those features were masked. In Figure 13, the same network misclassified a riot image from London as a non-riot after masking. This indicates that Score-CAM was able to identify the features that drove classification. The regions of interest in both Figures 12 and 13 appear to be similar and have similar ULU makeups, but the impact of masking had different results. This is not offered as an explanation for why the network struggled with images from Russia yet performed well in England. But this does highlight the challenges of attempting to explain the features that drive classification for satellite images of conflict and the strong potential of deep learning techniques to identify features that may not be evident to human interpreters.

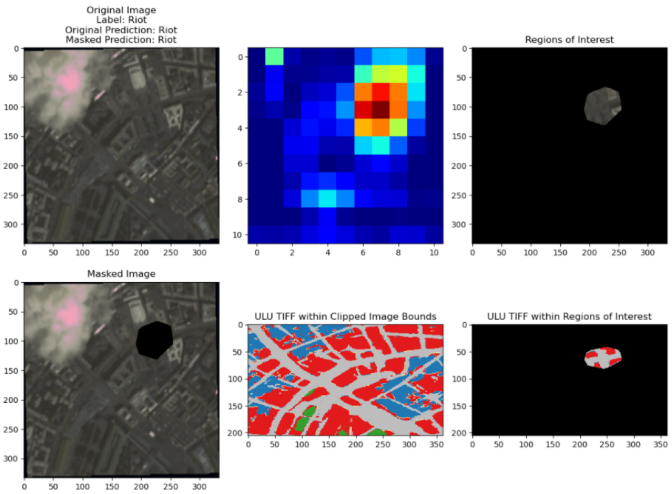

Figure 12. An image from a riot in Moscow. This figure displays the same Score-CAM, masking, and ULU methodology used in the rest of this research. In this example, a riot image from central Moscow was properly classified as a riot. Following masking, the image was still classified as a riot. This indicates that Score-CAM was not able to properly identify the location of the features that drove riot classification. Imagery ©Planet Labs PBC 2023. All rights reserved.

There is no clear relationship between ResNet's classification accuracy and Score-CAM's ability to identify relevant features. If Score-CAM consistently identified riot features, we would expect a uniform performance degradation across countries, regardless of the initial accuracy. Alternatively, if Score-CAM's performance correlated with ResNet's accuracy, we would expect larger accuracy drops in countries where ResNet performed well and smaller drops where it struggled. The results show neither pattern. For instance, Yemen and Peru, which had lower unmasked accuracies, exhibited significant performance drops after masking, while Russia showed a minimal impact despite similar initial accuracy levels.

The differences in the Score-CAM performance may have stemmed from several factors. There is no evidence to suggest that riot-related features are universal or consistent across countries. Cultural or regional factors could influence the visual indicators of riots, making features in one country easier to detect and isolate than in another. Additionally, Score-CAM may be well suited to identifying certain types of features while struggling with others, depending on their complexity, context, or spatial relationships within the imagery.

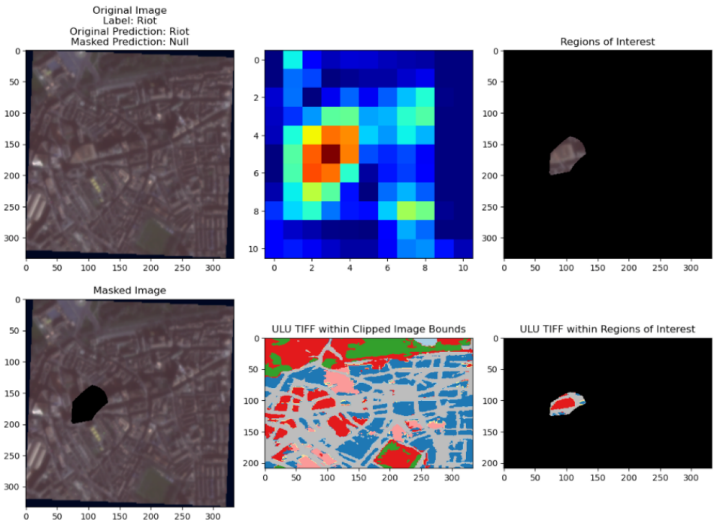

Figure 13. An image from a riot in London. This figure displays the same Score-CAM, masking, and ULU methodology used in the rest of this research. In this example, a riot image from London was correctly classified as a riot prior to masking. After masking, the image was incorrectly classified as a non-riot. This indicates that Score-CAM was able to identify the location of features in the image that drove riot classification. Imagery ©Planet Labs PBC 2023. All rights reserved.

These variations highlight the need for further investigation into the applicability and limitations of explainable AI techniques like Score-CAM for satellite imagery. All of these findings are predicated on using ResNet18 and Score-CAM, and further exploration of different networks and techniques is warranted. Understanding these challenges is critical for improving feature detection and interpretation across diverse geographic and socioeconomic contexts, making this an important area for future research.

5. Conclusions

In this paper, we sought to answer two key questions: first, *how effective are class activation maps at localizing the regions of satellite imagery that are relevant for the classification of socioeconomic factors*, and second, *can the land use be used as a proximate factor to understand the likely location of riot events in urban environments?* Through our analysis, we established that Score-CAM was not able to consistently identify the regions of the image critical to the neural network's classification decisions, demonstrating mixed utility as an interpretative tool in high-stakes, complex applications like conflict prediction. Furthermore, our comparison with urban land use (ULU) data revealed that these identified features did not consistently correlate with traditional land cover classifications, indicating that the neural network learned to recognize unique visual patterns associated with riot-prone areas, potentially capturing subtle socioeconomic signals that would not be conventionally detected. The observed lack of a correlation holds significant implications for advancing research on the integration of deep learning with traditional GIS techniques, suggesting that prediction or explainability predicated on the land cover may be insufficient in some cases.

These findings carry significant implications for the use of explainable AI in the satellite-based predictive modeling of complex social phenomena. By moving beyond

traditional land cover explanations, this study opens up new avenues for interpreting deep learning models applied to satellite imagery in contexts where specific correlates are unknown or semantically ill defined. Challenges remain, particularly regarding the interpretability of features that do not map easily onto known semantic definitions (i.e., land cover types or objects such as cars or households). Future work may benefit from integrating additional data sources to refine these interpretations, ultimately advancing both the robustness and transparency of AI-driven models for social and urban research. Future research in this field should extend beyond the independent evaluation of the model performance and interpretability, instead emphasizing the development of models that achieve high performance while inherently providing interpretability. This approach presents a dual challenge, as it requires balancing the the predictive accuracy with transparency, a task that may be intrinsically complex given the nature of satellite imagery, where critical features of importance are often unknown or not explicitly defined.

Author Contributions: Conceptualization, S.W. and D.R.; methodology, S.W. and D.R.; validation, S.W. and D.R.; formal analysis, S.W. and D.R.; investigation, S.W. and D.R.; resources, D.R.; data curation, S.W. and D.R.; writing—original draft preparation, S.W. and D.R.; writing—review and editing, S.W. and D.R.; visualization, S.W. and D.R.; supervision, D.R.; project administration, D.R.; funding acquisition, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation, Award Number 2317591. The data were made available through the NASA Commercial Smallsat Data Acquisition (CSDA) program. This work was funded by the U.S. Department of Homeland Security and the Center for Accelerating Operating Efficiency, Arizona State University, Grant Award Number 17STQAC00001-03-03.

Data Availability Statement: The data that support the findings of this study are available from Planet Labs PBC. Restrictions apply to the availability of these data, which were used under license for this study. The data are available from the author with the permission of Planet Labs PBC.

Acknowledgments: The authors acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACLED	Armed Conflict Location and Event Data Project
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
DEGURB	Degree of Urbanisation
GIS	Geographic Information System
LIME	Local Interpretable Model-Agnostic Explanations
LULC	Land Use and Land Cover
RGB	Red, Green, and Blue
ROC	Receiver Operating Characteristic
Score-CAM	Score-Weighted Class Activation Mapping
SHAP	Shapley Additive Explanations
ULU	Urban Land Use
XAI	Explainable Artificial Intelligence

Appendix A

Appendix A.1

As a part of the analysis of any country-specific findings in Section 3.4, the similarity scores, average softmax shifts, and the distribution of softmax shifts for each country were explored. The following charts are presented to provide more insight into the similar behavior and performance across all countries in the dataset. Figures A1–A3 present the L1, L2, and Cosine similarities for each country. Further softmax shift analyses are shown in Figures A5 and A6.

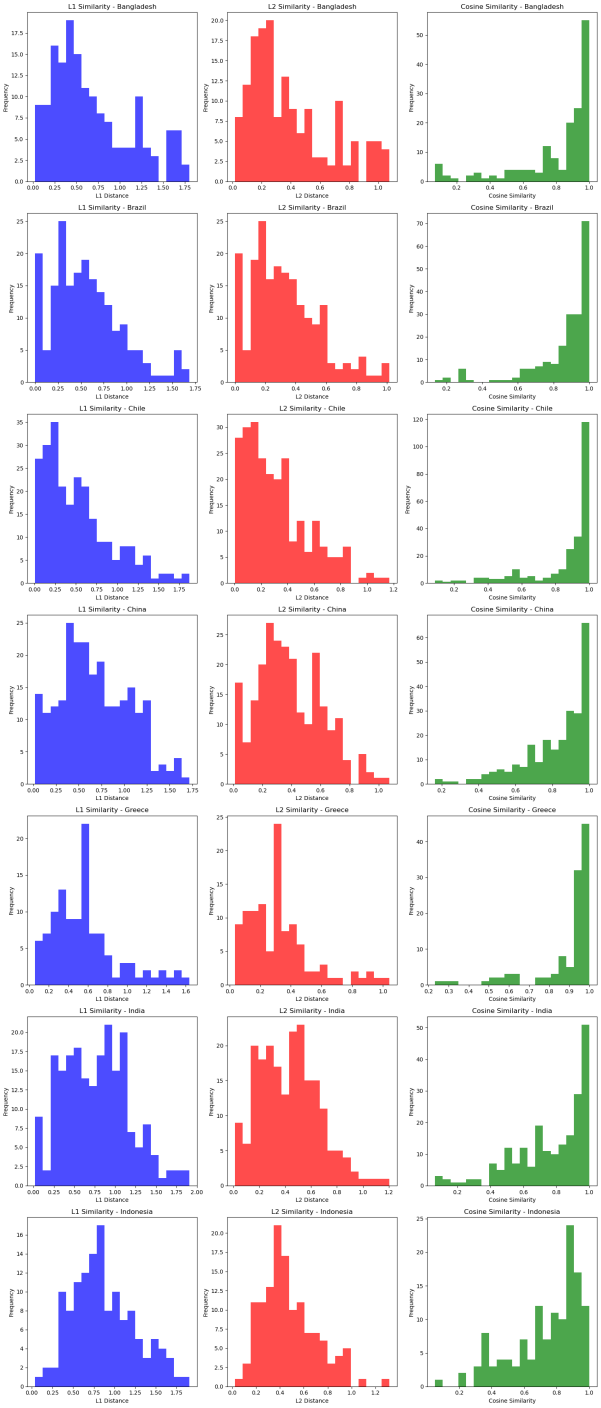

Figure A1. Country-specific similarities. The frequency of L1 similarities is shown in blue. The frequency of L2 similarities is shown in red. The frequency of Cosine similarities is shown in green.

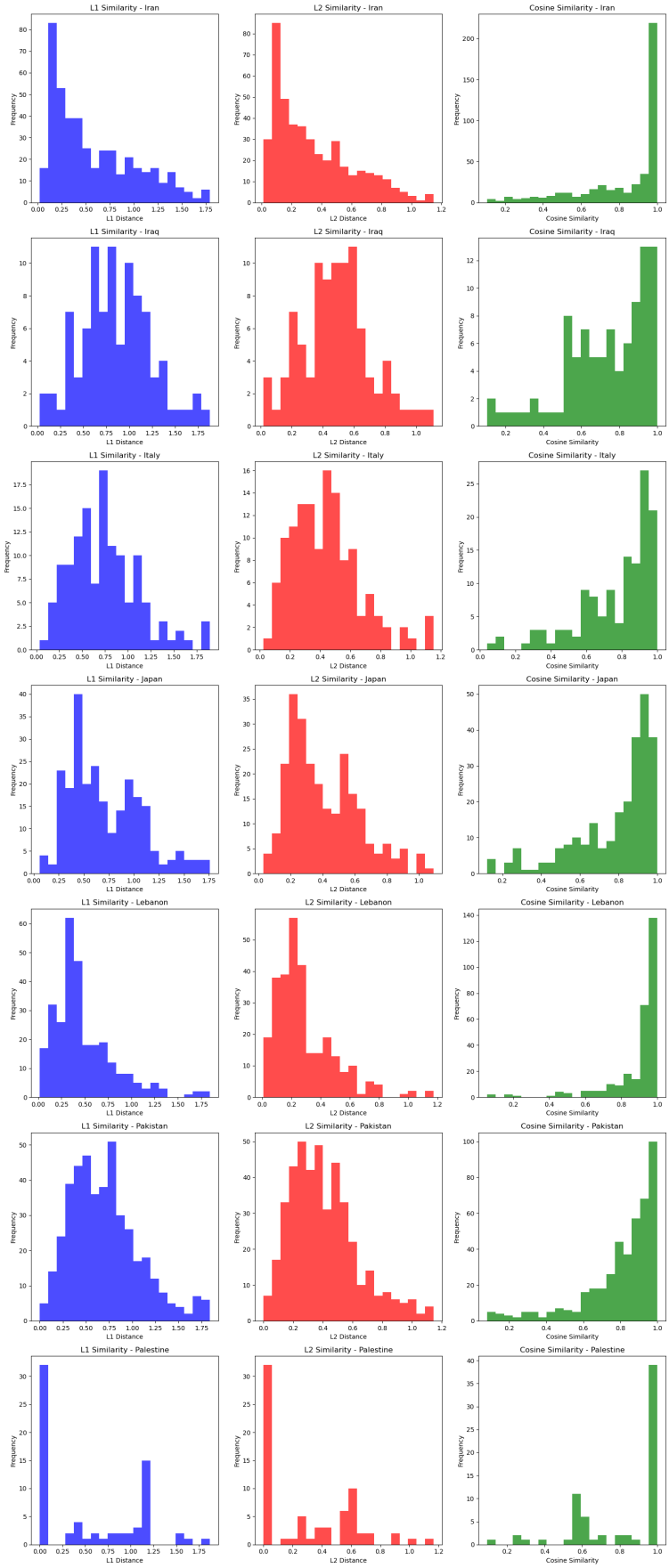

Figure A2. Country-specific similarities. The frequency of L1 similarities is shown in blue. The frequency of L2 similarities is shown in red. The frequency of Cosine similarities is shown in green.

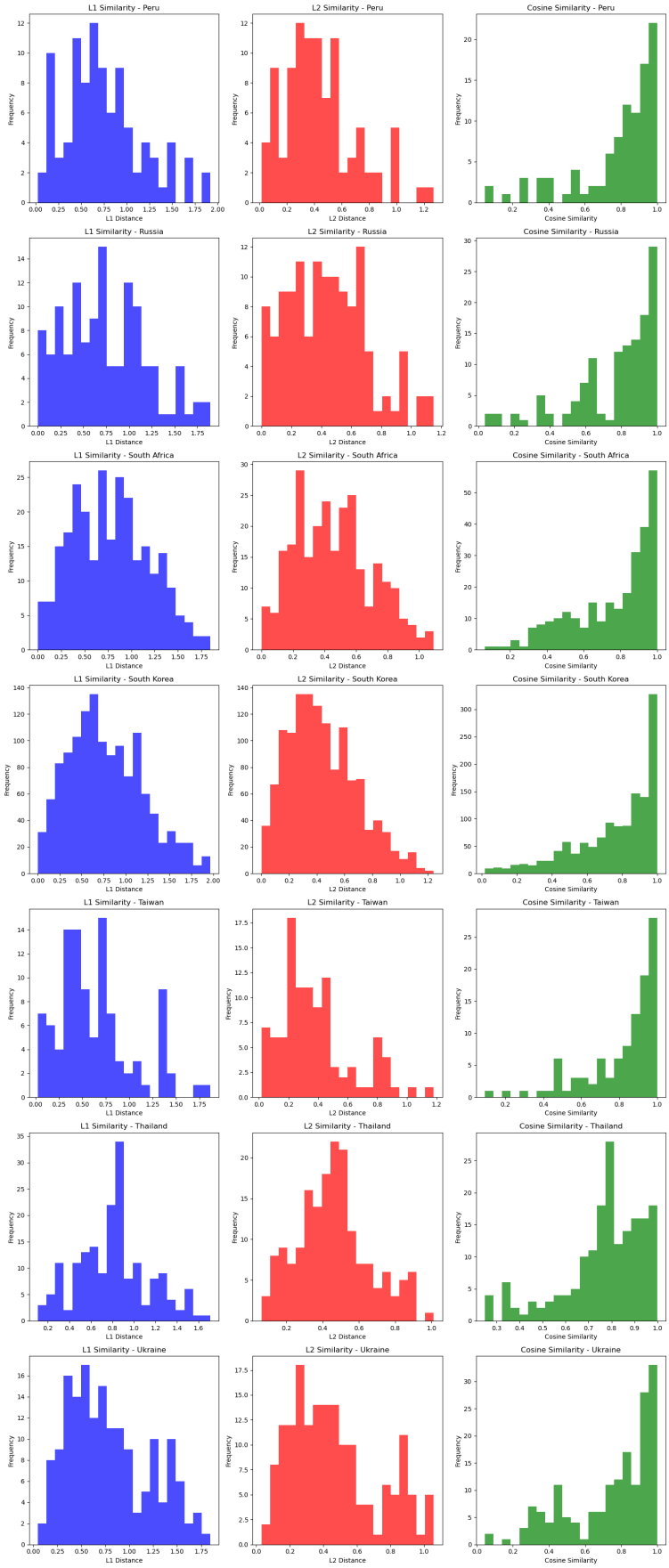

Figure A3. Country-specific similarities. The frequency of L1 similarities is shown in blue. The frequency of L2 similarities is shown in red. The frequency of Cosine similarities is shown in green.

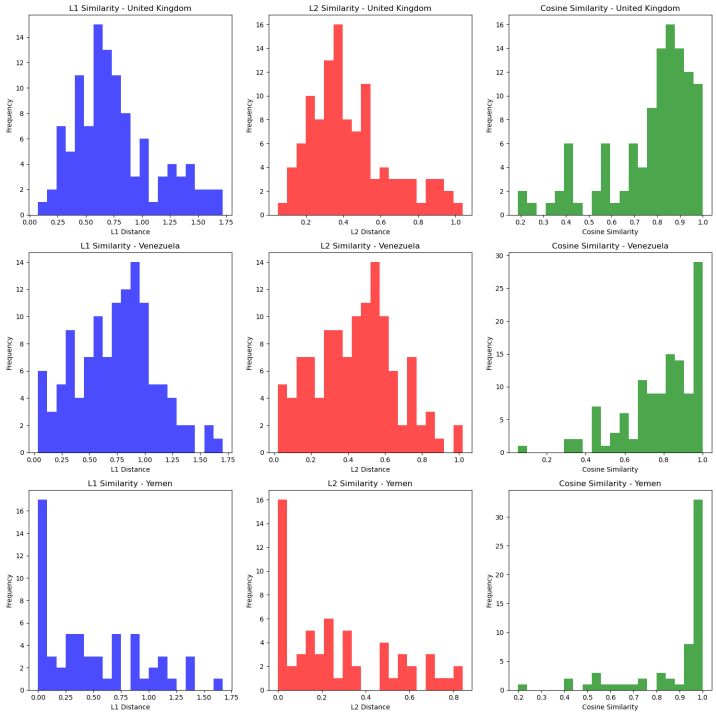

Figure A4. Country-specific similarities. The frequency of L1 similarities is shown in blue. The frequency of L2 similarities is shown in red. The frequency of Cosine similarities is shown in green.

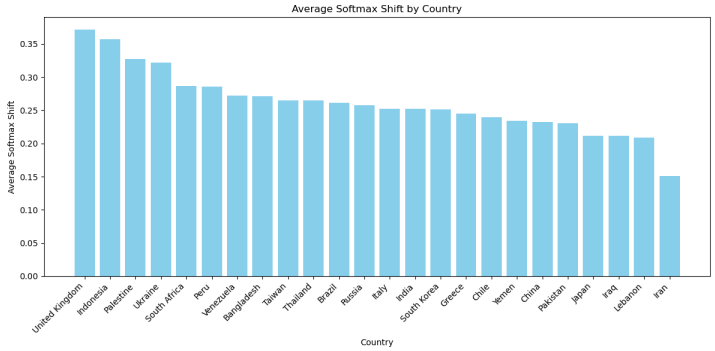

Figure A5. The country-specific softmax shift. This is a bar graph of the mean for each country.

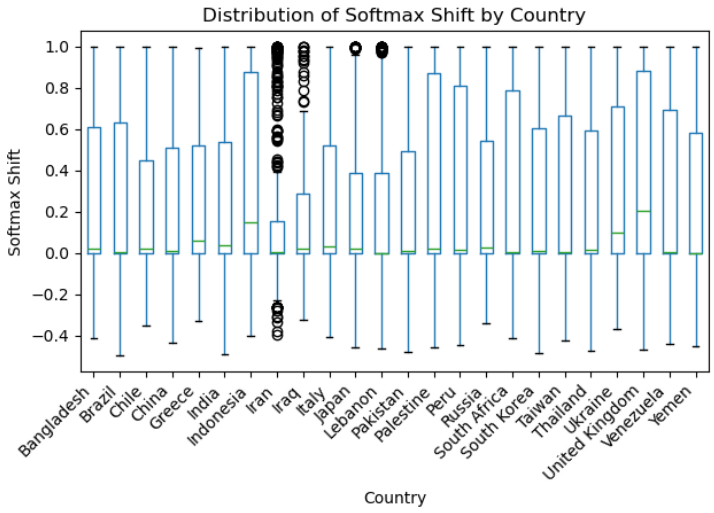

Figure A6. The country-specific softmax shift. This is a box plot displaying the distribution of the softmax shift for each country.

References

1. Rodgers, D.; Gazdar, H.; Goodfellow, T. *Cities and Conflict*; London School of Economics and Political Science (LSE): London, UK, 2010.
2. Askarizad, R.; Safari, H. The influence of social interactions on the behavioral patterns of the people in urban spaces (case study: The pedestrian zone of Rasht Municipality Square, Iran). *Cities* **2020**, *101*, 102687. [CrossRef]
3. Snow, D.A.; Vliegenthart, R.; Corrigall-Brown, C. Framing the French riots: A comparative study of frame variation. *Soc. Forces* **2007**, *86*, 385–415. [CrossRef]
4. Davies, T.P.; Fry, H.M.; Wilson, A.G.; Bishop, S.R. A mathematical model of the London riots and their policing. *Sci. Rep.* **2013**, *3*, 1303. [CrossRef]
5. Warnke, S.; Runfola, D. Predicting Protests and Riots in Urban Environments With Satellite Imagery and Deep Learning. *Trans. GIS* **2024**, *28*, 2309–2327. [CrossRef]
6. Jean, N.; Burke, M.; Xie, M.; Davis, W.M.; Lobell, D.B.; Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science* **2016**, *353*, 790–794. [CrossRef]
7. Runfola, D.; Stefanidis, A.; Lv, Z.; O'Brien, J.; Baier, H. A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 726–750. [CrossRef]
8. Runfola, D.; Stefanidis, A.; Baier, H. Using satellite data and deep learning to estimate educational outcomes in data-sparse environments. *Remote Sens. Lett.* **2022**, *13*, 87–97. [CrossRef]
9. Runfola, D.; Baier, H.; Mills, L.; Naughton-Rockwell, M.; Stefanidis, A. Deep learning fusion of satellite and social information to estimate human migratory flows. *Trans. GIS* **2022**, *26*, 2495–2518. [CrossRef]
10. Goodman, S.; BenYishay, A.; Runfola, D. A convolutional neural network approach to predict non-permissive environments from moderate-resolution imagery. *Trans. GIS* **2021**, *25*, 674–691. [CrossRef]
11. Aung, T.S.; Overland, I.; Vakulchuk, R.; Xie, Y. Using satellite data and machine learning to study conflict-induced environmental and socioeconomic destruction in data-poor conflict areas: The case of the Rakhine conflict. *Environ. Res. Commun.* **2021**, *3*, 025005. [CrossRef]
12. Goodman, S.; BenYishay, A.; Runfola, D. Spatiotemporal Prediction of Conflict Fatality Risk Using Convolutional Neural Networks and Satellite Imagery. *Remote Sens.* **2024**, *16*, 3411. [CrossRef]
13. Planet Team. *Planet Application Program Interface: In Space for Life on Earth*; Digital Globe: San Francisco, CA, USA, 2023.
14. Obadic, I.; Levering, A.; Pennig, L.; Oliveira, D.; Marcos, D.; Zhu, X. Contrastive Pretraining for Visual Concept Explanations of Socioeconomic Outcomes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 575–584.
15. Machicao, J.; Specht, A.; Vellenich, D.; Meneguzzi, L.; David, R.; Stall, S.; Ferraz, K.; Mabile, L.; O'brien, M.; Corrêa, P. A deep-learning method for the prediction of socio-economic indicators from street-view imagery using a case study from Brazil. *Data Sci. J.* **2022**, *21*, 1929464. [CrossRef]
16. Bansal, C.; Jain, A.; Barwaria, P.; Choudhary, A.; Singh, A.; Gupta, A.; Seth, A. Temporal prediction of socio-economic indicators using satellite imagery. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. Association for Computing Machinery, Hyderabad, India, 5–7 January 2020; pp. 73–81.
17. Hall, O.; Ohlsson, M.; Rögnvaldsson, T. A review of explainable AI in the satellite data, deep machine learning, and human poverty domain. *Patterns* **2022**, *3*, 100600. [CrossRef]
18. Dabkowski, P.; Gal, Y. Real time image saliency for black box classifiers. *arXiv* **2017**, arXiv:1705.07857.
19. Fong, R.C.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3429–3437.
20. Petsiuk, V.; Das, A.; Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv* **2018**, arXiv:1806.07421.
21. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: New York, NY, USA, 2018; pp. 839–847.
22. Naidu, R.; Ghosh, A.; Maurya, Y.; Nayak K, S.R.; Kundu, S.S. IS-CAM: Integrated Score-CAM for axiomatic-based explanations. *arXiv* **2020**, arXiv:2010.03023.
23. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part I 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
24. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
25. Dosovitskiy, A.; Brox, T. Inverting visual representations with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837.

26. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
27. Vaswani, A. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
28. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
30. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL Tech.* **2017**, *31*, 841. [CrossRef]
31. Höhl, A.; Obadic, I.; Torres, M.Á.F.; Najjar, H.; Oliveira, D.; Akata, Z.; Dengel, A.; Zhu, X.X. Opening the Black-Box: A Systematic Review on Explainable AI in Remote Sensing. *arXiv* **2024**, arXiv:2402.13791.
32. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
33. Yamauchi, T.; Ishikawa, M. Spatial sensitive grad-cam: Visual explanations for object detection by incorporating spatial sensitivity. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: New York, NY, USA, 2022; pp. 256–260.
34. Sattarzadeh, S.; Sudhakar, M.; Plataniotis, K.N.; Jang, J.; Jeong, Y.; Kim, H. Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 1775–1779.
35. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 24–25.
36. Shi, T.; Li, Y.; Liang, H.; Yu, R. Score-CAMpp: Class activation map based on logarithmic transformation. In Proceedings of the 2022 16th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 21–24 October 2022; IEEE: New York, NY, USA, 2022; Volume 1, pp. 256–259.
37. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
38. Vasu, B.; Rahman, F.U.; Savakis, A. Aerial-cam: Salient structures and textures in network class activation maps of aerial imagery. In Proceedings of the 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Aristi Village, Zagori, Greece, 10–12 June 2018; IEEE: New York, NY, USA, 2018; pp. 1–5.
39. Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images. *Remote Sens.* **2019**, *11*, 544. [CrossRef]
40. Simonyan, K. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
41. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
42. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
43. Yang, F.; Xu, Q.; Li, B. Ship detection from optical satellite images based on saliency segmentation and structure-LBP feature. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 602–606. [CrossRef]
44. Temenos, A.; Temenos, N.; Kaselimi, M.; Doulamis, A.; Doulamis, N. Interpretable deep learning framework for land use and land cover classification in remote sensing using SHAP. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 8500105. [CrossRef]
45. Khan, M.; Hanan, A.; Kenzhebay, M.; Gazzea, M.; Arghandeh, R. Transformer-based land use and land cover classification with explainability using satellite imagery. *Sci. Rep.* **2024**, *14*, 16744. [CrossRef]
46. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv* **2020**, arXiv:2009.07896.
47. Tahir, A.; Munawar, H.S.; Akram, J.; Adil, M.; Ali, S.; Kouzani, A.Z.; Mahmud, M.P. Automatic target detection from satellite imagery using machine learning. *Sensors* **2022**, *22*, 1147. [CrossRef] [PubMed]
48. Carleer, A.; Debeir, O.; Wolff, E. Assessment of very high spatial resolution satellite image segmentations. *Photogramm. Eng. Remote Sens.* **2005**, *71*, 1285–1294. [CrossRef]
49. Brewer, E.; Lin, J.; Runfola, D. Susceptibility & defense of satellite image-trained convolutional networks to backdoor attacks. *Inf. Sci.* **2022**, *603*, 244–261.

50. Burka, B.M.; Roro, A.G.; Regasa, D.T. Dynamics of pastoral conflicts in eastern Rift Valley of Ethiopia: Contested boundaries, state projects and small arms. *Pastoralism* **2023**, *13*, 5. [CrossRef]
51. Tan, S.; Hassen, N.A. Examining the choice of land conflict resolution mechanisms: The case between the harshin and yocaale woredas of the Somali region of Ethiopia. *J. Environ. Manag.* **2023**, *342*, 118250. [CrossRef] [PubMed]
52. Kugler, T.A.; Grace, K.; Wrathall, D.J.; de Sherbinin, A.; Van Riper, D.; Aubrecht, C.; Comer, D.; Adamo, S.B.; Cervone, G.; Engstrom, R.; et al. People and Pixels 20 years later: The current data landscape and research trends blending population and environmental data. *Popul. Environ.* **2019**, *41*, 209–234. [CrossRef]
53. Council, N.R.; on Environmental Change, B.; on the Human Dimensions of Global Change, C. *People and Pixels: Linking Remote Sensing and Social Science*; National Academies Press: Cambridge, MA, USA, 1998.
54. Seto, K.C.; Güneralp, B.; Hutyra, L.R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16083–16088. [CrossRef] [PubMed]
55. Walsh, S.J.; Crews-Meyer, K.A.; Crawford, T.W.; Welsh, W.F. Population and environment interactions: Spatial considerations in landscape characterization and modeling. *Scale Geogr. Inq. Nature Soc. Method* **2004**, 41–65. [CrossRef]
56. Rindfuss, R.R.; Walsh, S.J.; Turner, B.L.; Fox, J.; Mishra, V. Developing a science of land change: Challenges and methodological issues. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13976–13981. [CrossRef] [PubMed]
57. Runfola, D.S.M.; Pontius Jr, R.G. Measuring the temporal instability of land change using the Flow matrix. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 1696–1716. [CrossRef] [PubMed]
58. Fortier, J.; Rogan, J.; Woodcock, C.E.; Runfola, D.M. Utilizing temporally invariant calibration sites to classify multiple dates and types of satellite imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 181–189. [CrossRef]
59. Alo, C.A.; Pontius Jr, R.G. Identifying systematic land-cover transitions using remote sensing and GIS: The fate of forests inside and outside protected areas of Southwestern Ghana. *Environ. Plan. B Plan. Des.* **2008**, *35*, 280–295. [CrossRef]
60. Stow, D.; Hamada, Y.; Coulter, L.; Anguelova, Z. Monitoring shrubland habitat changes through object-based change identification with airborne multispectral imagery. *Remote Sens. Environ.* **2008**, *112*, 1051–1061. [CrossRef]
61. Rogan, J.; Chen, D. Remote sensing technology for mapping and monitoring land-cover and land-use change. *Prog. Plan.* **2004**, *61*, 301–325. [CrossRef]
62. Li, X.; Chen, D.; Duan, Y.; Ji, H.; Zhang, L.; Chai, Q.; Hu, X. Understanding Land use/Land cover dynamics and impacts of human activities in the Mekong Delta over the last 40 years. *Glob. Ecol. Conserv.* **2020**, *22*, e00991. [CrossRef]
63. Murillo-Sandoval, P.J.; Kilbride, J.; Tellman, E.; Wrathall, D.; Van Den Hoek, J.; Kennedy, R.E. The post-conflict expansion of coca farming and illicit cattle ranching in Colombia. *Sci. Rep.* **2023**, *13*, 1965. [CrossRef] [PubMed]
64. Zhang, J.; Niu, J.; Buyantuev, A.; Wu, J. A multilevel analysis of effects of land use policy on land-cover change and local land use decisions. *J. Arid. Environ.* **2014**, *108*, 19–28. [CrossRef]
65. Addae, B.; Oppelt, N. Land-use/land-cover change analysis and urban growth modelling in the Greater Accra Metropolitan Area (GAMA), Ghana. *Urban Sci.* **2019**, *3*, 26. [CrossRef]
66. Wu, Y.; Li, S.; Yu, S. Monitoring urban expansion and its effects on land use and land cover changes in Guangzhou city, China. *Environ. Monit. Assess.* **2016**, *188*, 54. [CrossRef] [PubMed]
67. Mandal, J.; Ghosh, N.; Mukhopadhyay, A. Urban growth dynamics and changing land-use land-cover of megacity Kolkata and its environs. *J. Indian Soc. Remote Sens.* **2019**, *47*, 1707–1725. [CrossRef]
68. Ishtiaque, A.; Shrestha, M.; Chhetri, N. Rapid urban growth in the Kathmandu Valley, Nepal: Monitoring land use land cover dynamics of a himalayan city with landsat imageries. *Environments* **2017**, *4*, 72. [CrossRef]
69. Runfola, D.M.; Hughes, S. What makes green cities unique? Examining the economic and political characteristics of the grey-to-green continuum. *Land* **2014**, *3*, 131–147. [CrossRef] [PubMed]
70. Runfola, D.M.; Polsky, C.; Nicolson, C.; Giner, N.M.; Pontius Jr, R.G.; Krahe, J.; Decatur, A. A growing concern? Examining the influence of lawn size on residential water use in suburban Boston, MA, USA. *Landsc. Urban Plan.* **2013**, *119*, 113–123. [CrossRef]
71. Yin, J.; Dong, J.; Hamm, N.A.; Li, Z.; Wang, J.; Xing, H.; Fu, P. Integrating remote sensing and geospatial big data for urban land use mapping: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102514. [CrossRef]
72. Chen, B.; Xu, B.; Gong, P. Mapping essential urban land use categories (EULUC) using geospatial big data: Progress, challenges, and opportunities. *Big Earth Data* **2021**, *5*, 410–441. [CrossRef]
73. Duncan, M.J.; Winkler, E.; Sugiyama, T.; Cerin, E.; Dutoit, L.; Leslie, E.; Owen, N. Relationships of land use mix with walking for transport: Do land uses and geographical scale matter? *J. Urban Health* **2010**, *87*, 782–795. [CrossRef] [PubMed]
74. Ewing, R.; Cervero, R. Travel and the built environment: A meta-analysis. *J. Am. Plan. Assoc.* **2010**, *76*, 265–294. [CrossRef]
75. Jacobs-Crisioni, C.; Rietveld, P.; Koomen, E.; Tranos, E. Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: An exploratory study using mobile phone data. *Environ. Plan. A* **2014**, *46*, 2769–2785. [CrossRef]
76. Frank, L.D.; Schmid, T.L.; Sallis, J.F.; Chapman, J.; Saelens, B.E. Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. *Am. J. Prev. Med.* **2005**, *28*, 117–125. [CrossRef]

77. Jia, Y.; Ge, Y.; Ling, F.; Guo, X.; Wang, J.; Wang, L.; Chen, Y.; Li, X. Urban land use mapping by combining remote sensing imagery and mobile phone positioning data. *Remote Sens.* **2018**, *10*, 446. [CrossRef]
78. Guan, D.; Li, H.; Inohae, T.; Su, W.; Nagaie, T.; Hokao, K. Modeling urban land use change by the integration of cellular automaton and Markov model. *Ecol. Model.* **2011**, *222*, 3761–3772. [CrossRef]
79. Verburg, P.H.; Schot, P.P.; Dijst, M.J.; Veldkamp, A. Land use change modelling: Current practice and research priorities. *GeoJournal* **2004**, *61*, 309–324. [CrossRef]
80. Batty, M. *The New Science of Cities*; MIT Press: Cambridge, MA, USA, 2013.
81. Alberti, M.; Marzluff, J.; Hunt, V.M. Urban driven phenotypic changes: Empirical observations and theoretical implications for eco-evolutionary feedback. *Philos. Trans. R. Soc. B Biol. Sci.* **2017**, *372*, 20160029. [CrossRef]
82. Xu, J.Z.; Lu, W.; Li, Z.; Khaitan, P.; Zaytseva, V. Building damage detection in satellite imagery using convolutional neural networks. *arXiv* **2019**, arXiv:1910.06444.
83. Mueller, H.; Groeger, A.; Hersh, J.; Matranga, A.; Serrat, J. Monitoring war destruction from space using machine learning. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2025400118. [CrossRef]
84. Nabiee, S.; Harding, M.; Hersh, J.; Bagherzadeh, N. Hybrid U-Net: Semantic segmentation of high-resolution satellite images to detect war destruction. *Mach. Learn. Appl.* **2022**, *9*, 100381. [CrossRef]
85. Eklund, L.; Degerald, M.; Brandt, M.; Prishchepov, A.V.; Pilesjö, P. How conflict affects land use: Agricultural activity in areas seized by the Islamic State. *Environ. Res. Lett.* **2017**, *12*, 054004. [CrossRef]
86. Planet Team. *PlanetScope: Constellation and Sensor Overview*; Digital Globe: San Francisco, CA, USA, 2023.
87. Raleigh, C.; Kishi, R.; Linke, A. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 74. [CrossRef]
88. Schiavina, M.; Melchiorri, M.; Pesaresi, M. *GHS-SMOD R2023A—GHS Settlement Layers, Application of the Degree of Urbanisation Methodology (Stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, Multitemporal (1975–2030)*; European Commission, Joint Research Centre (JRC): Brussels, Belgium, 2023. [CrossRef]
89. European Commission and Statistical Office of the European Union. *Applying the Degree of Urbanisation—A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons—2021 Edition*; Publications Office of the European Union: Brussels, Belgium, 2021; ISBN 978-92-76-20306-3. [CrossRef]
90. Runfola, D.; Anderson, A.; Baier, H.; Crittenden, M.; Dowker, E.; Fuhrig, S.; Goodman, S.; Grimsley, G.; Layko, R.; Melville, G.; et al. geoBoundaries: A global database of political administrative boundaries. *PLoS ONE* **2020**, *15*, e0231866. [CrossRef] [PubMed]
91. Guzder-Williams, B.; Mackres, E.; Angel, S.; Blei, A.M.; Lamson-Hall, P. Intra-urban land use maps for a global sample of cities from Sentinel-2 satellite imagery and computer vision. *Comput. Environ. Urban Syst.* **2023**, *100*, 101917. [CrossRef]
92. ACLED Codebook. 2023. Available online: https://acleddata.com/acleddatanew/wp-content/uploads/dlm_uploads/2023/06/ACLED_Codebook_2023.pdf (accessed on 14 January 2025).
93. Pearce, T.; Brintrup, A.; Zhu, J. Understanding softmax confidence and uncertainty. *arXiv* **2021**, arXiv:2106.04972.
94. Subramanya, A.; Srinivas, S.; Babu, R.V. Confidence estimation in deep neural networks via density modelling. *arXiv* **2017**, arXiv:1707.07013.
95. Moon, J.; Kim, J.; Shin, Y.; Hwang, S. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*; PMLR: Birmingham, UK, 2020; pp. 7034–7044.
96. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* **2016**, arXiv:1610.02136.
97. Rozsa, A.; Günther, M.; Boult, T.E. Adversarial robustness: Softmax versus openmax. *arXiv* **2017**, arXiv:1708.01697.
98. Sen, J.; Sen, A.; Chatterjee, A. Adversarial Attacks on Image Classification Models: Analysis and Defense. *arXiv* **2023**, arXiv:2312.16880.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.