5-2023

Seeing What We Can't: Evaluating implicit biases in deep learning satellite imagery models trained for poverty prediction

Joseph O'Brien
William & Mary

WILLIAM & MARY
CHARTERED 1693

THE COLLEGE OF WILLIAM & MARY

HONORS THESIS

Seeing What We Can't: Evaluating implicit biases in deep learning satellite imagery models trained for poverty prediction

---

Author:
Joseph O'BRIEN

Advisor:
Dan RUNFOLA

A thesis submitted in fulfillment of the requirements for Interdisciplinary Honors in the degree of Bachelors of Science in the

Data Science Program

Accepted for Honors

Chair: Dr. Dan Runfola

Dr. Matthew Haug

Dr. Jaime Settle

Williamsburg, Virginia May 10, 2023

THE COLLEGE OF WILLIAM & MARY

HONORS THESIS

Seeing What We Can't: Evaluating implicit biases in deep learning satellite imagery models trained for poverty prediction

---

Author:
Joseph O'BRIEN

Advisor:
Dan RUNFOLA

A thesis submitted in fulfillment of the requirements for Interdisciplinary Honors in the degree of Bachelors of Science in the

Data Science Program

Accepted for Honors

Chair: Dr. Dan Runfola

Dr. Matthew Haug

Dr. Jaime Settle

Williamsburg, Virginia May 10, 2023

THE COLLEGE OF WILLIAM & MARY

Abstract

Dr. Dan Runfola
Data Science Program

Bachelors of Science

Seeing What We Can't: Evaluating implicit biases in deep learning satellite imagery models trained for poverty prediction

by Joseph O'BRIEN

Previous studies have sought to use Convolutional Neural Networks for regional estimation of poverty levels. However, there is limited research into possible implicit biases in deep neural networks in the context of satellite imagery. In this work, we develop a deep learning model to predict the tertile of per-capita asset consumption, trained on satellite imagery and World Bank Living Standards Measurements Study data. Using satellite imagery collected via survey location data as inputs, we use transfer learning to train a VGG-16 Convolutional Neural Network to classify images based on per-capita consumption. The model achieves an R^2 of .74, using thousands of observations across Ethiopia, Malawi, and Nigeria. Using a variety of interpretability techniques, our study seeks to qualitatively analyze images to evaluate implicit biases in the model. Our results indicate that roads, urban infrastructure, and coastlines are the three human-interpretable features that have the largest influence on the predicted consumption level for a given image.

Contents

List of Figures

List of Abbreviations

CNN	Convolutional Neural Network
ML	Machine Learning
FAT ML	Fair and Transparent Machine Learning
GAN	Generative Adversarial Network
LASSO	least absolute shrinkage and selection operator
SE	squeeze and excitation module
NPP-VIIRS	National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite
DNB	Day/Night Band
NTL	Nighttime Lights
VGG	Visual Geometry Group, University of Oxford
WI	Wealth Index
NDVI	Normalized Difference Vegetation Index
OSM	OpenStreetMap
LMICs	Low and Middle Income Countries
GDP	Gross Domestic Product
TRSCG	total retail sales of consumer goods
FPN	Feature Pyramid Netword
PCGDP	Per-capita Gross Domestic Product
GBPE	General Public Budgetary Expenditure
PCLERR	Per-capita living expenditure of rural residents
PCDIRR	Per-capita disposable income of rural residents
DHS	Demographic and Health Survey
SHAP	Shapley Additive Explanations
LSMS	Living Standards Measurements Study
PPP	Purchasing Power Parity
ReLU	Rectified Linear Unit Activation Function
SGD	Stochastic Gradient Descent

Chapter 1

Thesis

1 Introduction

As deep convolutional modeling has gained traction as a viable data-driven decision-making tool, these models have begun to see real-world use in sectors including the allocation of poverty relief aid (Jean et al., 2016) and population estimation in remote regions (Robinson, Hohman, and Dilkina, 2017). Taking advantage of neural network technologies can be challenging, as there are trade-offs between model accuracy and interpretability (Lipton, 2018). This is predominantly due to the very large number of parameters required to fit a neural network, resulting in a labor-intensive process to understand why a model makes a given estimate or recommendation.

Deep neural networks have been applied in a variety of contexts. A review of the applications of CNNs shows their ability for the application to a variety of problems, including home loan default prediction (Sirignano, Sadhwani, and Giesecke, 2016), spam detection (Crawford et al., 2015), classification of plant diseases (Hasan, Yusuf, and Alzubaidi, 2020), and medical imagery analysis (Ker et al., 2017). Recent work surveying interpretability in deep learning has sought to disambiguate the terms "interpretability" and "explainability" (Chakraborty et al., 2017). The authors propose that interpretability refers to the interpretation arrived at by an agent when given an output explanation from a model.

Jean et al., 2016 is widely considered to be the first research showing that convolutional neural networks (CNNs) are capable of accurately predicting poverty using satellite imagery data. However, there is limited research on the potential implicit biases in these models. Numerous studies have raised concerns about

the potential for biases due to factors ranging from model generalizability to the lack of Fair and Transparent Machine Learning (FAT ML) guidelines currently used in the development of CNNs (Burke et al., 2021; Hall, Ohlsson, and Rögnvaldsson, 2022). However, there still exists a lack of systematic analysis of model biases in poverty prediction using CNNs.

In this work, we seek to fill a gap in the existing literature exploring model interpretability and implicit biases in the context of deep learning techniques leveraging satellite imagery. We aim to answer the question: *Are there implicit biases in deep learning models that seek to identify impoverished populations using satellite imagery?*

To address the potential biases of CNNs in poverty prediction, we leverage interpretability techniques such as Backpropagation (Simonyan, Vedaldi, and Zisserman, 2014) and Grad-CAM (Selvaraju et al., 2017) to analyze model bias. These techniques allow us to identify which parts of the image might be most important in classifying images, and to visualize the neural network's decision-making process. By using interpretability techniques to analyze the model's outputs, we can identify potential biases in the model that may arise from image quality, data coverage, human interpretable features, and other factors. To the authors' knowledge, this is the first example of a systematic study of interpretability techniques being applied in the context of satellite imagery.

2 Literature Review

2.1 Utilizing Satellite Imagery in Deep Learning and Poverty Estimation

Recent work in transfer learning has shown that convolutional neural networks (CNNs) can predict poverty using high-resolution satellite imagery with a high level of accuracy (Jean et al., 2016), (Xie et al., 2016). In the seminal pieces on this topic, data on nighttime light intensity is used as a proxy for economic development. This data was used to train a CNN which predicted nighttime lights based on daytime imagery. Jean et al., 2016 showed that CNN approaches could explain up to 75% of the variation in local-level economic outcomes in sub-Saharan

Africa. Building on this work, Xie et al., 2016 showed that the CNN learned filters for man-made structures including roads, buildings, and farmlands. The authors write that these features are highly informative for predicting and mapping poverty and argue that their data approaches the accuracy of field-collected data. These papers are widely considered to be the first works on convolutional approaches to poverty prediction and they have proven especially important for estimating poverty in data-sparse areas where it may not be feasible to conduct on-the-ground surveys to estimate consumption and other economic variables.

Building on the seminal work of Jean et al., 2016 and Xie et al., 2016, Ayush et al., 2020 showed that work combining object recognition and regression analysis might be useful in predicting local-level poverty even in cases where the overall number of labels is small. The authors argue that their framework provides a high level of accuracy while providing policymakers with semantically meaningful features, hopefully leading to increased adoption of these techniques for predicting poverty and distributing aid. In studies comparing model architectures and imagery resolutions, researchers found that a GoogleNet model trained on Digital Globe imagery performed slightly better than other approaches in explaining poverty variance in 896 Mexico municipalities (Babenko et al., 2017). Researchers concluded that model accuracy did not carry over to municipalities outside the 896 included in the 2014 Módulo de Condiciones Socioeconómicas-Encuesta Nacional de Ingresos y Gastos de los Hogares (MCS-ENIGH) survey (a survey of socioeconomic conditions in Mexico). Researchers hypothesize that this lack of generalizability could be due to weighting geographic tiles by land area rather than population or that MCS-ENIGH municipalities have characteristics different from non-MCS-ENIGH municipalities. The authors conclude that more work is needed to determine the ways in which training processes influence sample validation.

Jean et al., 2019 propose Tile2Vec, an unsupervised representation learning algorithm that uses theories from linguistics and applies them to geospatial information. Using the distributional hypothesis, the idea that words that appear in similar contexts typically have similar meanings, the authors argue that spatial features may also co-occur, and the vectorization of such information might allow researchers to better predict poverty and other economic indicators. The Tile2Vec approach was able to explain 49.6% of the variance in poverty data in

Uganda as opposed to transfer learning approaches which explained 41% of the variance (Jean et al., 2016). These gains were achieved even while using imagery with a lower resolution than the transfer learning approach.

Perez et al., 2019 attempted to recreate the results of Jean et al. (Jean et al., 2016) using a generative adversarial network (GAN) approach. The GAN approach allows the model to be used to simultaneously predict the output of several semi-supervised tasks, giving it potential for use in situations where labeled data is scarce. The discriminator GAN trained by the authors achieved a validation accuracy of 65% when using all nine imagery bands and a validation accuracy of 68% when using only the red, green, and blue bands.

Ni et al., 2020 tested four deep learning architectures (VGG-Net, InceptionNet, ResNet, and DenseNet) to extract features from daytime satellite imagery, using least absolute shrinkage and selection operator (LASSO) regression to predict poverty. The authors integrate a squeeze and excitation (SE) module in the CNNs which is used to model channel relationships and interdependencies and include a form of self-attention. Results from this work showed that the SE module and focal loss into DenseNet performed the best.

Recent work has noted the limitations of our current deep learning approaches. In reviewing state-of-the-art approaches the deep learning poverty prediction, Jarry et al., 2021 concluded that spatial perturbation injected in the coordinates of poverty indicators significantly reduces the prediction power of the models. A grid-cell selection method showed improved performance in solving the spatial perturbation issue and researchers suggested further developing this work in order to continue to improve performance.

Traditionally, poverty prediction using CNNs has relied on nighttime light intensity as a proxy for economic development (Jean et al., 2016;Liu et al., 2021 write that gradient visualization showed their CNN was capable of detecting visual patterns thought to be closely related to economic development and confirmed the validity of using nighttime lights as a proxy for economic development. The team trained a VGG-16 model on the 2017 county-level GDP in mainland China and predicted 2018 values with an R^2 of 0.71. Yeh et al., 2020 used a CNN to predict survey-based poverty estimates for over 20,000 villages throughout Africa. The team showed that the model explained 70% of the variation in ground-measured socioeconomic status in countries where the model

was not trained, outperforming previous benchmarks. Satellite-based estimates were also able to account for 50% of the variation in change in wealth over time. The authors note that while their model outperforms simpler models more common to the literature, there is likely a hesitancy for these models to be adopted by the policy community, due to the lack of interpretability in deep learning approaches. The authors suggest that future work should aim to develop approaches to navigate apparent performance-interpretability tradeoffs and that these approaches are best used to augment current on-the-ground survey efforts.

Zhao et al., 2019 used a Vgg16 model trained on data from the National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) Day/Night Band (DNB) nighttime light (NTL) data, Google satellite imagery, land cover map, road map, and location data for the capital of each division in Bangladesh. The extracted features were used to train a random forest regression which successfully predicted the household wealth index (WI) with an R^2 value of 0.70 in Bangladesh and 0.61 in Nepal. The researchers also identified that proximity to urban areas was the most important variable in explaining poverty, contributing 37.9% of the explanatory power. Further contributions to the literature have been made by teams who showed that these techniques generalize to areas in Mexico (Babenko et al., 2017, the Philippines and Thailand (Hofer et al., 2020), India (Daoud et al., 2021), and Brazil (Castro and Álvarez, 2023. While training CNNs on satellite imagery and nighttime lights has proven effective in predicting relative levels of poverty, it is still an open question whether these methods can predict changes in poverty over time. Researchers found that these methods struggled to quantify changes in the economic status of communities in Rwanda between 2005 and 2015 (Kondmann and Zhu, 2020).

Research teams have begun to use other strategies in predicting poverty. Work building on Xie et al., 2016 demonstrated that models trained on images of varying resolutions were capable of extracting semantically different features than single-resolution models (Kim et al., 2016). The authors write that these multi-resolution models were able to outperform previous models trained with only a single resolution and that models trained on different resolutions extract "semantically different features that are often conceivable by human sight."

Work by researchers in India introduced a two-step approach to using CNNs to predict poverty in rural India (Pandey, Agarwal, and Krishnan, 2018). The team first trained a multi-task CNN to predict various developmental parameters from roof material to water and light sources. The network learned other meaningful features like roads and water bodies. Then a model was trained using the developmental parameter outputs of the first model to predict the income levels. Tang, Liu, and Matteson, 2022 showed that consumption levels can be predicted using normalized difference vegetation index (NDVI) rather than solely nighttime lights. Researchers hypothesize that NDVI is effective in predicting economic outcomes in rural areas which are highly dependent on agriculture. NDVI was also effective in predicting consumption variation over time. Chen, 2017 studied ways in which data aggregation might affect model performance. The team aggregated data by time and location and explored how aggregating all available data would positively impact predictive power. The team concluded that grouping imagery temporally and then training the network performed the best of the tested methods. They also found that while major increases in the amount of data did not hurt validation loss or accuracy, it did not lead to increased correlation. Other researchers (Kondmann and Zhu, 2020) suggest incorporating data fusion or refining CNN architectures to predict changes in local activity directly.

Recent work has begun to look for ways to leverage open-source data, recognizing the prohibitive costs associated with downloading imagery and training deep learning models. Building on the work done by Jean et al., 2016, researchers showed that models trained using lower-resolution, publicly available Landsat 7 imagery can achieve accuracy that exceeds previous benchmarks (Perez et al., 2017). This increase in performance while using lower-resolution imagery was likely due to the inclusion of hyperspectral imagery bands which are unavailable in Google Static Maps. This research showed that while CNN-based approaches are accurate in predicting relative wealth in countries where ground-truthed data is available, they may struggle to generalize and predict economic outcomes across country boundaries. Other work focused on using publicly available crowd-sourced geospatial information to implement the studies from Xie et al., 2016 and Jean et al., 2016 in the Philippines (Tingzon et al.,

2019). The research team was able to achieve an r-squared value of 0.63 for estimating asset-based wealth and showed that models trained on a combination of OpenStreetMap (OSM) and nighttime light-derived features are effective for real-time poverty mapping. Ayush et al., 2021 proposed a novel reinforcement learning algorithm that aims to decrease the number of high-resolution imagery tiles needed to accurately predict consumption and other economic outcomes. High-resolution imagery is often delivered by private companies, and as such, can be prohibitively expensive for teams operating on smaller budgets. The proposed algorithm decreases the necessary amount of high-resolution imagery by 80% while retaining the accuracy of models which exhaustively use high-resolution imagery.

2.1.1 Using CNNs to Create Global & Real Time Poverty Maps

Early work by Tingzon et al., 2019 showed the potential for using CNNs to create real-time poverty maps. Building on this work, Chi et al., 2022 used deep learning to develop micro-estimates of relative poverty at a 2.4-kilometer resolution for all 134 low and middle-income countries (LMICs). The researchers suggest that these maps could be useful in evaluating the impacts of the COVID-19 pandemic or other events with potential socioeconomic impacts. They also note that maps have a tendency to reproduce historical inequities and that due to the technically complicated and expensive processes of explaining model predictions, there is an opportunity for data manipulation and misreporting if predictions are not produced or validated by independent bodies. Other work created high-resolution poverty maps at a 1 square mile resolution for 25 sub-Saharan African countries (Lee and Braithwaite, 2022). The researchers then applied their modeling approach to the remaining 19 sub-Saharan African countries for which they lacked data. In work discussing the challenges of using deep learning to create poverty maps in Sierra Leone, researchers showed that model performance improved when noisy, ground-truthed clusters are moved closer to the nearest populated rural place (Espın-Noboa, Kertész, and Karsai, 2022). Their work also showed that population and mobility features were the strongest predictors of wealth. Researchers concluded that satellite images were best for predicting poor populations, while metadata features were best for predicting rural and urban-middle classes.

2.1.2 Predicting Other Human Wellness Indicators

Other work has sought to use CNNs trained on satellite imagery to estimate a broader range of socioeconomic indicators. Studies by Wu and Tan, 2019a showed that a Resnet50 model trained on nighttime lights was capable of predicting gross domestic product (GDP) and total retail sales of consumer goods (TRSCG) with a Pearson coefficient of 0.85. Wu and Tan, 2019b applied this work to areas of Guizhou Province, China, and found that these approaches were transferable to other regional areas. Tan et al., 2020 suggest combining a ResNet-50 and a feature pyramid network (FPN), to solve the multiscale problem in object detection. In experiments in Chongqing, China, researchers showed that this architecture could be trained to predict per-capita gross domestic product (PCGDP), general public budgetary expenditure (GPBE), per-capita living expenditure of rural residents (PCLERR), and per-capita disposable income of rural residents (PCDIRR).

Head et al., 2017 study showed that work done predicting poverty in sub-Saharan Africa could be generalized to other countries and continents, but model performance varied based on hyperparameter tuning. Researchers also concluded that these approaches did not generalize to other estimators of development in Africa, including educational attainment, access to drinking water, and a variety of health-related indicators. Work exploring the use of deep learning in predicting education outcomes showed that using satellite imagery and school test scores, accuracy for individual scores across years with an accuracy between 76% to 80% Runfola, Stefanidis, and Baier, 2022.

Irvin, Laird, and Rajpurkar, 2017 used the Demographic and Health Survey (DHS) data to predict health outcomes for data-sparse regions. Researchers noted that pre-trained models trained on smaller datasets performed better than a fully trained model and that the first layers of a ResNet model were capable of detecting infrastructure and human development. The researchers also concluded that mistakes made by the classification model typically were only one class off, showing that these models have difficulty distinguishing between the poor and middle ranges in the same ways that humans do. Models for malnutrition developed by the researchers failed to generalize well to the testing data set in all experiments. Researchers postulate that this may be due to noise in the malnutrition data and that very few direct features of malnutrition exist in

satellite imagery. Runfola et al., 2022 showed that using data fusion to combine census data with satellite imagery allowed CNNs to predict human migratory flows with an accuracy of R^2=0.72, an improvement over models leveraging only socioeconomic data by 10%.

Other researchers have provided the first evidence that deep learning approaches are capable of conducting the evaluation of anti-poverty programs using satellite imagery (Huang, Hsiang, and Gonzalez-Navarro, 2021). The team noted that one limitation of these models is that the impacts of anti-poverty programs must be observable from space, and this prevents the accurate evaluation of programs that do not directly impact the built environment. Another limitation is that while wealth is measured at the household or individual level, satellites capture features of villages or regions. This might make these approaches less effective in measuring the impact of programs that impact human mobility.

2.1.3 Impediments to Integrating Deep Learning Models into Decisionmaking

Reviews of work on convolutional and deep learning approaches to poverty estimation have noted that while these models show promising results, even operational models have failed to be fully embraced by decision-makers (Burke et al., 2021). The authors write that this is likely because decision and policymakers are unlikely to adopt measures they are unable to fully understand. The authors also write that while CNNs provide end users with a high level of predictive performance, they tend to do so while sacrificing interpretability. Burke et al. (Burke et al., 2021) cite improved model interpretabilty and transparency as one of the areas in which future work will be particularly useful. The authors suggest applying guidelines for Fairness, Accountability, and Transparency in Machine Learning (FAT ML) but write that they are currently unaware of any papers that have adequately engaged with these frameworks. In another review of deep learning techniques (Hall, Ohlsson, and Rögnvaldsson, 2022), the authors suggest that while recent papers have shown that these techniques can be made to generalize across several countries (Chi et al., 2022; Lee and Braithwaite, 2022), explainability is essential to the continuation of progress in the field and that explainability should be understood as broader than mere interpretability.

2.2 Frameworks for Model Interpretability

A review of the current literature suggests that there is currently no single accepted definition of interpretability in the machine-learning community. Lipton (Lipton, 2018) proposes two possibilities in response to this insight. It could be that the definition of interpretability is universally agreed upon, but no research paper has laid out that definition in writing. It might be that interpretability is not well defined, and thus, claims about interpretable models may only be "quasi-scientific" in nature. Lipton argues that a survey of research in the field of model interpretability suggests that the latter suggestion is correct.

Lipton suggests that interpretability is not a "monolithic concept but several distinct ideas that must be disentangled before any progress can be made." Lipton settles on five desiderata of interpretability: trust, causality, transferability, informativeness, and fair and ethical decision-making, and argues that current techniques to illuminate decision-making in models fall into two broad categories: transparency and post-hoc explanations. Transparency techniques involve describing how the model works at the level of the entire model, while post-hoc explanations are typically focused on providing the user with information extracted from the model. Lipton argues that model transparency and post-hoc explanations are competing concepts, with post-hoc explanations having the added advantage of being model agnostic. Post-hoc techniques are used after the prediction step and can thus be used on a variety of models without sacrificing the performance of these models.

Gilpin et al., 2022 writes that "while there are certainly technical problems with generating explanations of Machine Learning (ML) models, they are overshadowed by the problem of the lack of agreement as to what we mean by the term." The authors present a framework for evaluating model explanations and argue that a one size fits all approach to explainability is insufficient for delivering insights to users. They argue that model explanations must take into account their functional role, the intended audience, and the capabilities of the system coupled with its access to data. The authors write that when researchers design explanations, they must work to consider the role that they play in the overall use of the system.

Gilpin et al., 2022 also discuss the dangers of focusing on "getting users to

trust a system rather than on making a system trustworthy." While it is important for users to trust the systems they rely on, the authors suggest that fostering this trust is highly context-dependent, and relies on "trust of other experts, design decisions at the system and user experience levels, and ancillary components such as domain-informed conversational interfaces."

Miller, 2019 writes that explanations are "sought in response to particular counter-factual cases ... that is, people do not ask why event P happened, but rather why event P happened instead of some event Q." The author also argues that humans rarely expect an explanation to deliver the complete cause of the event and that while humans are capable of selecting infinitely many causes from an infinite set of causal explanations, these selections are likely influenced by our own cognitive biases. Miller, 2019 showed that humans are susceptible to placebic information and cognitive biases when evaluating explanations and causal links. Miller argues that future work in interpretability must leverage social science research if we wish to develop machine learning models and artificial intelligence that is worthy of trust.

2.3 Assessing Dataset Bias in Computer Vision

Vardi, 2022 discusses how generalization in neural networks might induce an implicit bias, and points to neural networks' vulnerability to adversarial examples as evidence for the implication of implicit bias in these systems. Many researchers have also focused on evaluating bias in the training datasets for neural networks (Khosla et al., 2012; Tommasi et al., 2017; Zhang, Wang, and Zhu, 2018; Deviyani, 2022).

Much of the work evaluating imagery dataset bias has been in the context of deep learning algorithms trained for facial recognition. Garvie and Frankle, 2016 write that while companies market facial recognition technologies as having an accuracy of over 95%, there likely exists racial bias in the training dataset, resulting in higher rates of misclassification for populations with darker skin tones. Ferguson, 2017 and Castro, 2019 explore facial recognition bias through the lens of policing specifically. Terhörst et al., 2021 explores bias in facial imagery datasets in areas beyond demographics. Researchers found that many non-demographic attributes strongly affect model performance and recognition

capabilities, such as accessories, hairstyle or hair color, face shapes, or facial anomalies.

To date, there has been no work published that directly assesses bias in satellite imagery datasets used to train CNNs. It is important to note, however, that because many CNNs leveraging satellite imagery are trained using transfer learning, they may still be susceptible to dataset bias in their original training.

2.4 Interpretability Techniques in Computer Vision

A variety of model-agnostic techniques have been developed for interpreting the outputs of CNNs. Ribeiro, Singh, and Guestrin, 2016 develop LIME, an explanation technique which they write "explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction." The authors build on this algorithm by developing SP-LIME, capable of selecting a set of representative explanations useful in building user trust in the model using submodular optimization. Ribeiro, Singh, and Guestrin, 2016 conclude that LIME users were better able to determine which models best generalized to the real world and were also able to improve the performance of classifiers by doing feature engineering using LIME. Lundberg and Lee, 2017 develop a coalitional game theoretic approach called SHAP (Shapley Additive Explanations), which can be used to explain the output of any machine learning model. This method uses Shapley values to assign each feature a value for the average marginal contribution to the overall prediction. The biggest difference between LIME and SHAP occurs in the weighting of instances in the model. While LIME weights instances according to how close they are from the original instance, SHAP weights instances according to the weight the coalition would receive in the Shapley value estimation.

Figure 1.1 shows an example of explanations output by LIME for the top two predicted classes for an image of rolls. In the middle image, LIME highlights the pixels which positively contribute to the classification of "bagel." The third image shows the explanation for the second most likely predicted class, "strawberry". In this image, we can see both pixels that positively contributed to this prediction in green and those that negatively influenced the prediction in red.

FIGURE 1.1: LIME explanations for the top 2 classes (bagel, strawberry) for image classification made by Google's Inception V3 neural network. (Image originally from Molnar, 2020)

Other researchers have worked to develop gradient-only methods which work by testing whether a change in pixel value would lead to a change in the predicted class of the image. Simonyan, Vedaldi, and Zisserman, 2014 presents two techniques for use in visualizing CNNs trained for image classification. The first outputs an artificially generated image which is representative of the class of interest. The second creates a class-specific saliency map, highlighting pixels in the images with respect to the specified class. This work demonstrated that understandable visualizations of CNN models could be produced using numerical optimization of input images. The researchers also showed that gradient-based visualization methods could generalize the deconvolutional network reconstruction procedure outlined by Zeiler and Fergus, 2014. Building on the work of Zhou et al., 2016, Selvaraju et al., 2017 develop Grad-CAM (Gradient-weighted Class Activation Mapping), which uses the gradients for a target classification to produce a coarse localization map highlighting the important regions in an image for the prediction of the target label. The authors write that Grad-CAM may also be useful for assessing dataset bias. Selvaraju et al., 2017 suggest that by using Grad-CAM to highlight areas of importance in the selected images, researchers can develop more holistically representative datasets by adding and removing images.

Recently, work has been done to study instances in which various post-hoc interpretability methods may disagree with one another. Krishna et al., 2022 performed analyses of six different interpretabilty methods to determine when

a disagreement arose and designed a qualitative user study to determine why and how researchers choose particular interpretability methods. In studying interpretability methods for imagery data, the researchers found that KernelSHAP (Lundberg and Lee, 2017) and LIME (Ribeiro, Singh, and Guestrin, 2016) had a higher agreement coefficient on all six of their metrics (rank correlation, pairwise rank agreement, feature agreement, rank agreement, sign agreement, and signed rank agreement) when compared to tabular and text data. However, their tests showed high levels of disagreement when computing rank correlation at the pixel level for gradient-based methods. The authors suggest that disagreement in interpretability methods for imagery varies significantly depending on image granularity. The authors conclude that work should be done to continue to educate scientists about novel approaches that can be used to resolve disagreements between explanations, and researchers should work to develop evaluation metrics that can be used to assist practitioners in evaluating which explanations are reliable in cases of disagreement.

3 Data

3.1 Study Area

Our study was designed to replicate - to the degree feasible under data availability constraints - the architecture developed by Jean et al., 2016. As such, Malawi and Nigeria were included as two of the countries for which imagery was downloaded. Ethiopia was also included in the study, despite not being included in the original Jean study, in order to explore the capability of the model in contexts outside of the original study (Mathur, 2020).

3.2 LSMS Clustering and Imagery Download

First, data from the World Bank's living standards measurements survey (LSMS) portal was downloaded for each country. Survey data from 2016-2017 was downloaded for Malawi, while 2015-2016 survey data was downloaded for Ethiopia and Nigeria. Using the LSMS data, clusters were generated to aggregate household information. Cluster coordinates were given by the LSMS data and are

defined as a 10km x 10km region enclosing a given central latitude and longitude. A total of 1967 cluster locations were provided in the survey data. Of these, 523 were in Ethiopia, 780 in Malawi, and 664 in Nigeria.

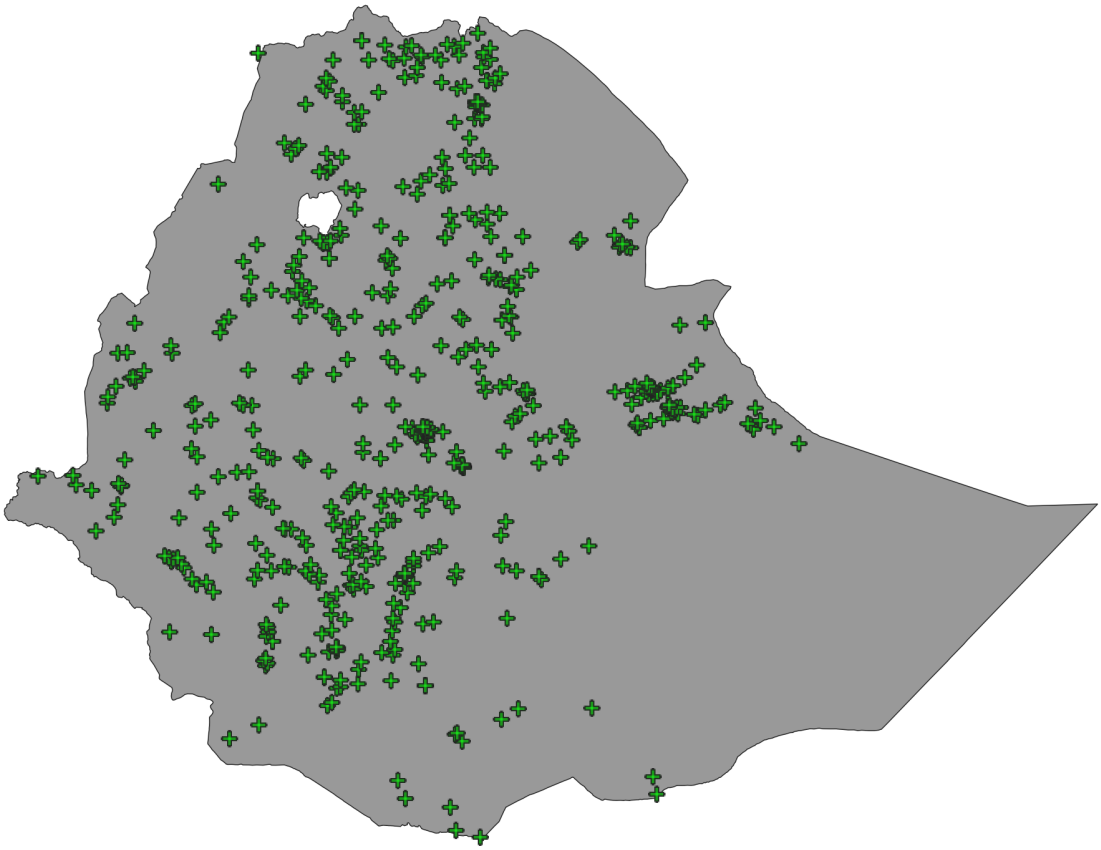

FIGURE 1.2: Cluster locations in Ethiopia

The LSMS data included information on household consumption, defined by the number of dollars spent on food per year. To get the household consumption per day, we use the information on the purchasing power parity (PPP) from the World Bank for each country for the year in which the LSMS data was collected. We then divide household income by PPP to standardize income and divide that value by 365 to get daily consumption by household (Where $1.90 is defined as the global poverty line). By adding the per-household consumption of all surveyed households within a cluster and dividing the aggregated consumption

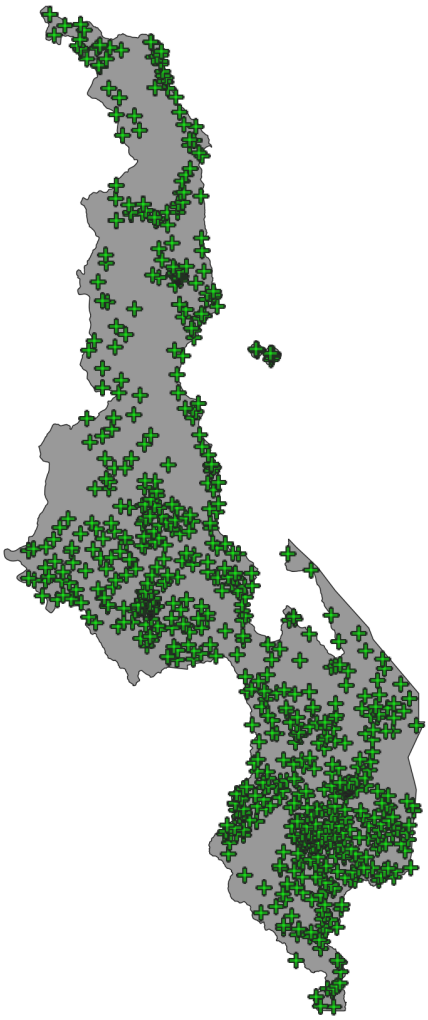

FIGURE 1.3: Cluster locations in Malawi

by the population in a given cluster we find the daily consumption per-capita. Per-capita consumption was then binned into tertiles and used as the image label to train the Vgg-16 model.

The imagery used to train the CNN was downloaded using the Planet API.[1] Images were filtered such that the most recent image taken between 2014-2016 with cloud cover of less than 5% was chosen for download. 8,501 images were downloaded in Ethiopia, 12,456 in Malawi, and 11,359 in Nigeria, totaling 32,316

[1]The Planet API was chosen over the Google Static Maps API as the Planet API allows users to filter imagery based on date, while the Google API only allows downloads of the most recent imagery. The imagery inside clusters was downloaded for dates ranging between January 2014 and December 2016.

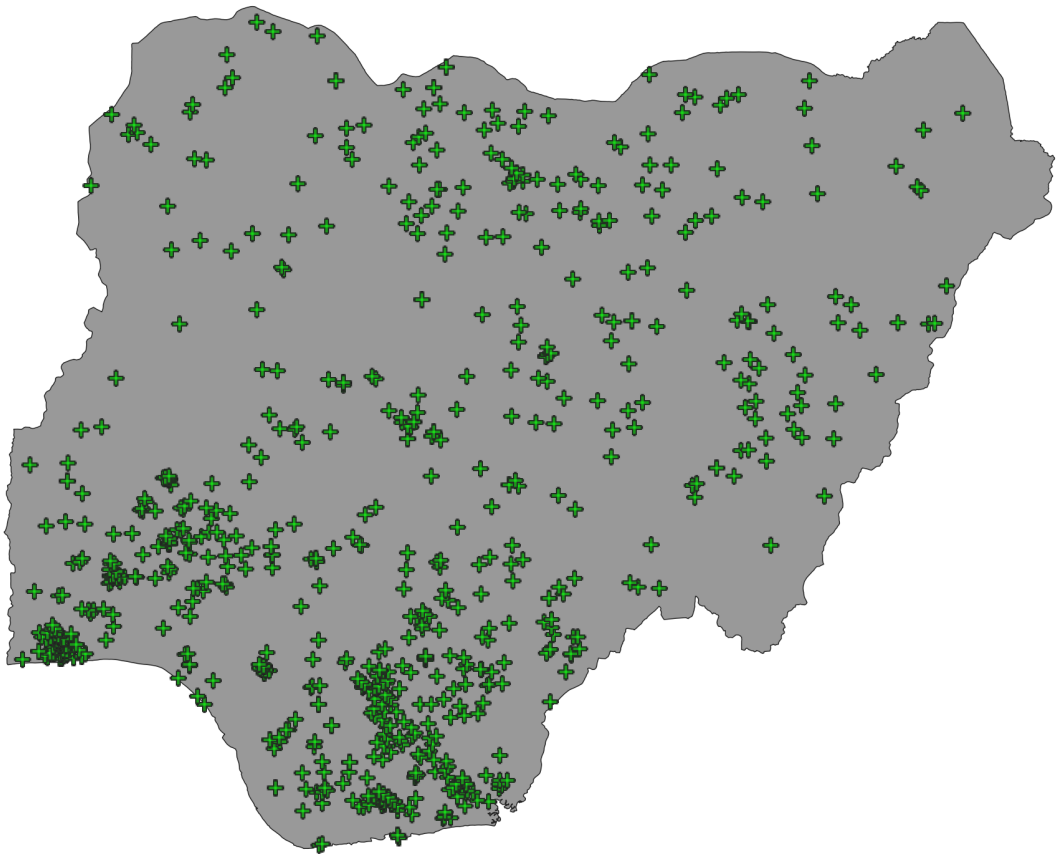

FIGURE 1.4: Cluster locations in Nigeria

images.

3.3 PlanetScope Imagery

Images of the PlanetScope Scene type from Planet are captured by the Dove Classic instrument. The data is collected using Planet's PSD telescope, which is capable of capturing red, green, blue, and near-infrared bands. The satellite produces imagery tiles which are approximately 25.0 x 11.5 sq km with a 3-meter pixel resolution. A total coverage area of 9,290,850 sq km was collected during the imagery download. This includes 2,444,037.5 sq km in Ethiopia, 3,581,100 sq km in Malawi, and 3,265,712.5 sq km in Nigeria.

The imagery downloaded was of the Basic Scene geometry type. As such, these images are not orthorectified or corrected for terrain distortions. The PlanetScope Basic Scene images are sensor-corrected and are scaled Top of Atmosphere Radiance (TOAR) at the sensor.

4 Methods

In our model, we seek to classify imagery into tertiles based on per-capita consumption. We do this based on satellite imagery alone, leveraging transfer learning on a Vgg-16 convolutional model (Tammina, 2019). The CNN created for this project was implemented using PyTorch, an open-source Python package.

4.1 Train and Test Split

The imagery was split into training and testing datasets within each model. Each of the clusters represents a 10km x 10km area surrounding a survey point from the LSMS data. All available imagery inside the cluster location was downloaded, resulting in nearly 25,000 images for 1,967 clusters. For each of the 1967 clusters, the training dataset received 80% of the imagery while the testing dataset received 20% of the imagery. For the training dataset, 12869 of the images were from Malawi, 8145 were from Ethiopia, and 4121 were from Nigeria. In the validation dataset, 3960 of the images were from Malawi, 2453 were from Ethiopia, and 1049 were from Nigeria. This represents imagery from the 523 clusters in Ethiopia, the 780 in Malawi, and the 664 in Nigeria.

4.2 Model Architecture

The Vgg-16 model (Simonyan and Zisserman, 2014) consists of 13 convolutional layers and three fully connected layers. A 3x3 filter is used in each of the convolutional layers, and after passing through a convolutional layer, the input is passed through the Rectified Linear Unit Activation Function (ReLU) which matches the output for positive inputs and outputs zero for negative inputs. Each group of convolutional layers is followed by a pooling layer which helps to reduce dimensionality. The final three layers are fully connected, and apply

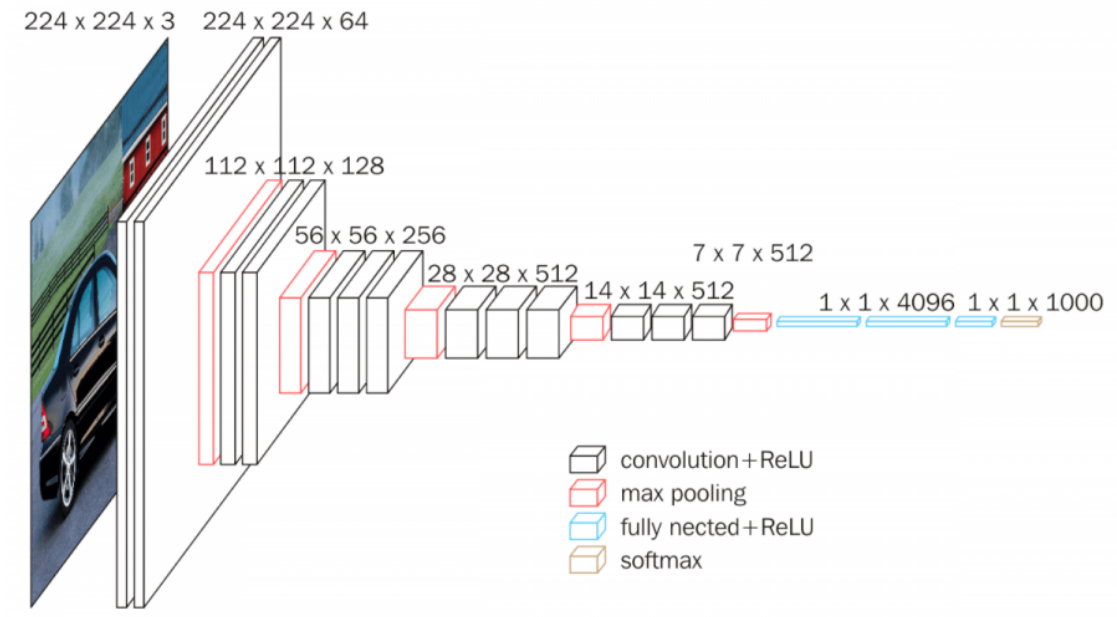

FIGURE 1.5: Network Structure of the Vgg-16 along with input dimensions.

a linear transformation to the input vector through a weights matrix to produce the final prediction class. Our model used stochastic gradient descent (SGD) with a momentum of 0.9 for optimization, a learning rate of .001, and a batch size of 8 to achieve a final validation accuracy of R^2=0.74.

4.3 Intepretability Methods

We use both Backpropagation for Saliency Mapping (Simonyan, Vedaldi, and Zisserman, 2014) and Grad-CAM (Selvaraju et al., 2017) as our two methods for assisting us in visualizing the neural network. In the case of Backpropagation, we get an output visualization highlighting the pixels of importance at the level of the whole model. In contrast, Grad-CAM produces a heat map of which pixels are most likely to change the classification of the image at each layer of the network. We use the outputs from these interpretability techniques to fill in a matrix of which human-interpretable features are highlighted in each image.

1.6 and 1.7 show a illustrative example for the workings of Grad-CAM. 1.7 shows the output for each image, based on three different interpretability methods. We use the vanilla and Grad-CAM methods referenced here, and the output

images assist end users in determining areas of the image the model focuses on for a given classification. As seen in these outputs, leveraging a variety of interpretability methods assists end users in determining which features are emphasized by the model.

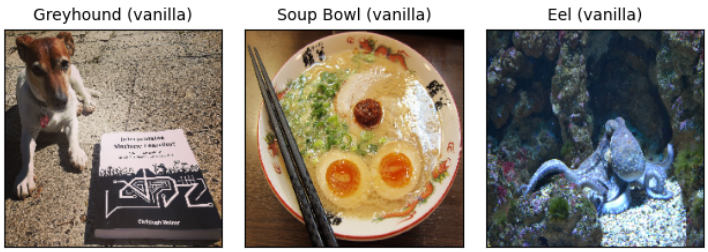

FIGURE 1.6: Original image for use in saliency mapping interpretability techniques

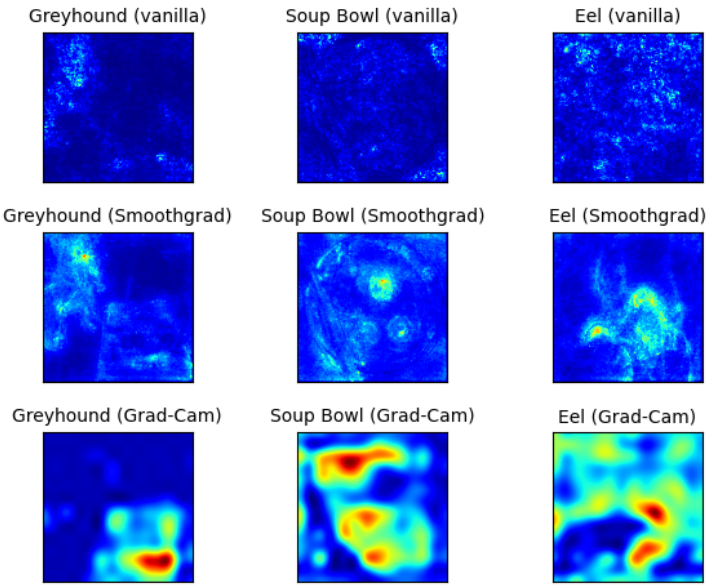

FIGURE 1.7: Saliency maps for three interpretability methods: Vanilla Gradient, SmoothGrad, and Grad-CAM

For our qualitative analysis, we selected 150 images (50 from each country) for testing that contained human-interpretable features on the ground. We identified nine categories of features that had the potential of being highlighted by our interpretability methods. These categories are river edges, coastlines, hillsides, croplands, human-built structures, straight roads, winding roads, road

intersections, and airports. For each of our test images, we pass the images through both interpretability methods and fill in the table with a one in the column if the feature was highlighted by interpretability methods and a zero if the feature was not highlighted or does not exist in the image [2].

5 Results

5.1 Base Model Results

Before exploring potential biases, in this section, we note how the results from our model are broadly similar to the results of the existing literature on using deep learning for poverty prediction. Previous work in the field falls into one of two broad categories for poverty prediction methodology. The first includes work by Jean et al., 2016, Chen, 2017, Xie et al., 2016, Perez et al., 2017, Tingzon et al., 2019, and others that seek to use extracted features from CNNs to train a regression model used to predict poverty. The second group includes work by Kim et al., 2016, Babenko et al., 2017, Head et al., 2017, Yeh et al., 2020, Ni et al., 2020, and others who aim to predict poverty directly using CNNs.

5.1.1 Model Calibration

Since we are most interested in the features that are extracted by the CNN, regardless of the use of these features in training regression models, we chose to train our CNN to classify images based on per-capita consumption. This provides us with a way to determine which human-interpretable features are extracted when predicting per-capita consumption directly, rather than training on nighttime lights as a proxy for socioeconomic development and training a regression model to predict per-capita consumption. Our model achieves a final validation accuracy of 0.74, which is comparable to the CNNs trained in other state-of-the-art research (0.75 (Jean et al., 2016);).71 (Xie et al., 2016; 0.61 Babenko et al., 2017). We achieved this accuracy using stochastic gradient descent (SGD) with a momentum of 0.9 for optimization, a learning rate of .001, and a batch size of 8.

[2] See appendix for the matrix of qualitative imagery analysis.

5.2 Manual Labeling & Qualitative Analysis

For each of our 150 selected images, we run them through both Backpropagation and Grad-CAM to produce output images that highlight the features emphasized in the prediction of the original imagery. For each of our nine classes of human-interpretable features (river edges, coastlines, hillsides, croplands, human-built structures, straight roads, winding roads, road intersections, and airports), we check our interpretability outputs to see if these features are emphasized by the model. In our matrix, we assign a value of 0 if the feature was not present or not highlighted and a value of 1 if a given feature was highlighted in either of our interpretability outputs[3]. Using this matrix, we construct a simple linear regression to identify which human-interpretable features were statistically significant for the prediction of per-capita consumption. After fitting the linear model, coastlines, winding roads, and buildings were identified as significant contributors to an accurate estimation of the level of per-capita consumption.

Of the 150 images that were qualitatively analyzed, 44 were misclassified by the CNN model. The misclassification rate of 29.3% is consistent with our model's validation accuracy of 0.74. In the cases where misclassification occurred, the observed class was two (the wealthiest households) in 17 cases, one (average households) in 19 cases, and zero (poorest households) in 8 cases. These map to per-capita consumption in the 66th-100th percentile for "wealthy", the 33rd-66th percentile for "average households", and the 0th-33rd percentile for "poor"

As a proportion of analyzed images, 25.75% of images in group two were misclassified, 37.25% of images in group one were misclassified, and 25.81% of images in group zero were misclassified. By country, 22% of images in Ethiopia were misclassified, 30% in Malawi, and 36% in Nigeria.

6 Discussion

Based on our qualitative analysis, we observe preliminary evidence that suggests CNN-based modeling approaches may fail in systematic ways which could

[3]See appendix for the matrix of qualitative imagery analysis.

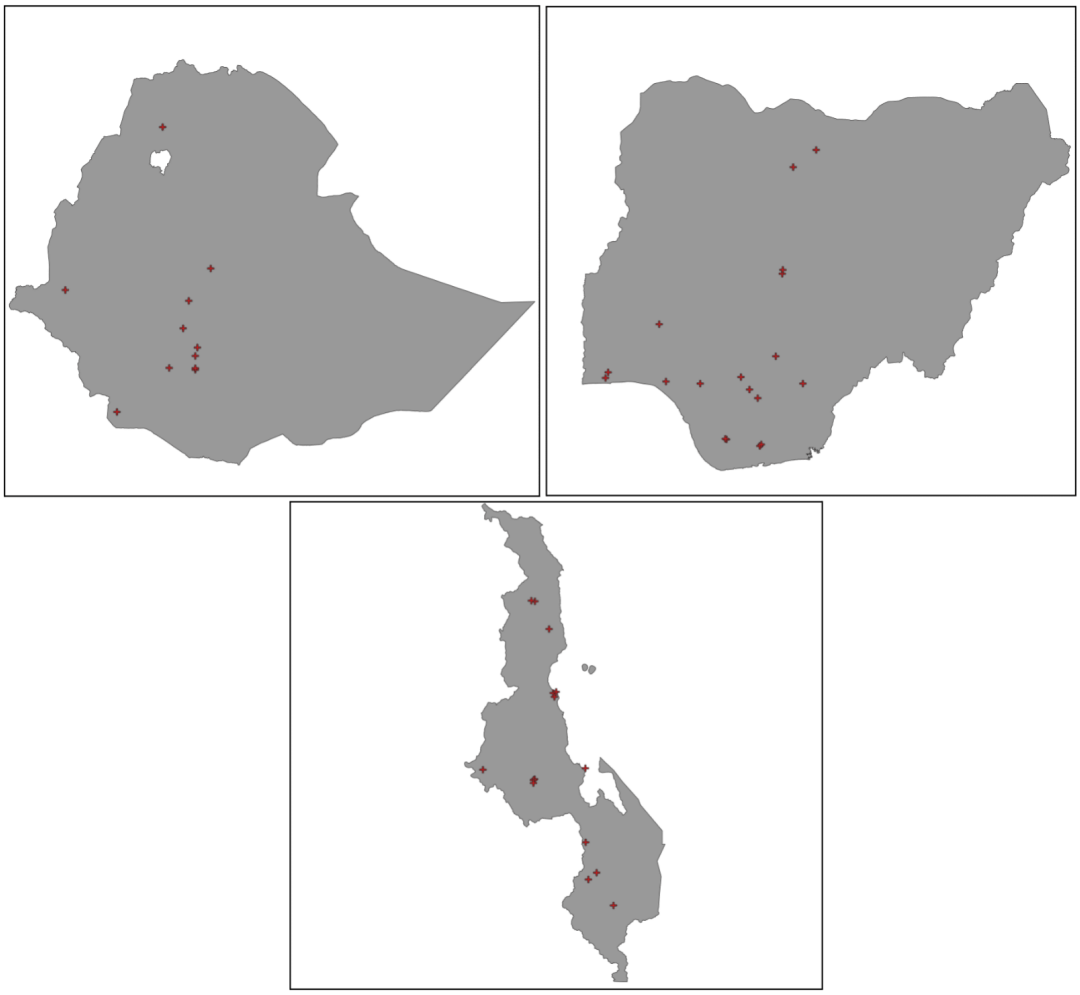

FIGURE 1.8: Locations of each of the misclassified images in each country.

lead to certain populations being disadvantaged by deep learning predictions of poverty. In this section, we explore the idea that these systems often fail at correctly classifying images in the same cases in which humans may, and sometimes fail in unique cases.

6.1 Classification Failures in Sparsely Populated Areas

It is interesting to note that only one of the qualitatively analyzed images was misclassified as being rich when the true classification was actually poor (i.e., off by two tertiles). While this is likely in part due to there only being three

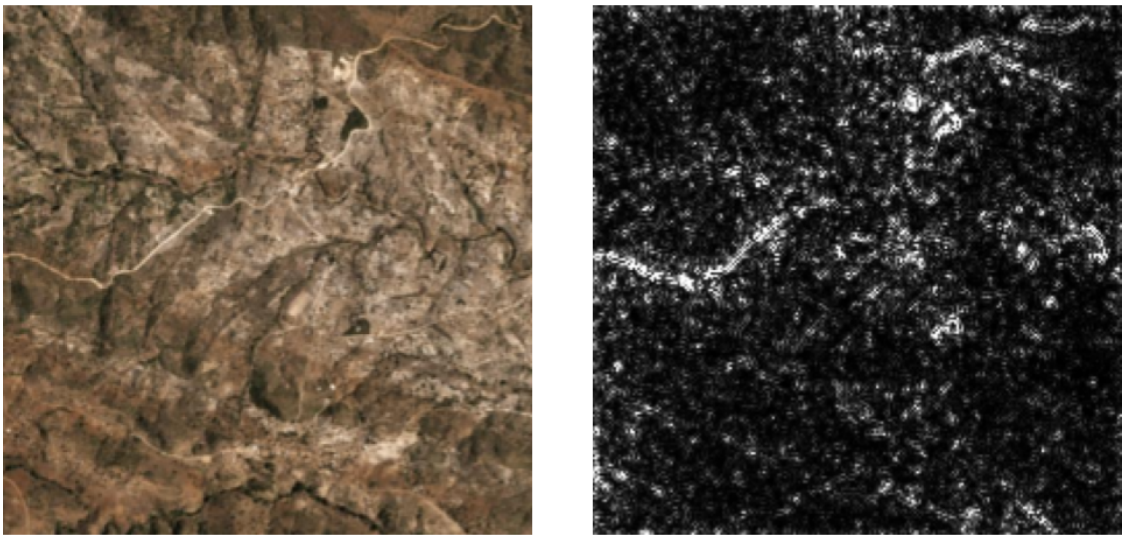

FIGURE 1.9: Original image and saliency map using backpropagation for the image at -14.919, 34.631.

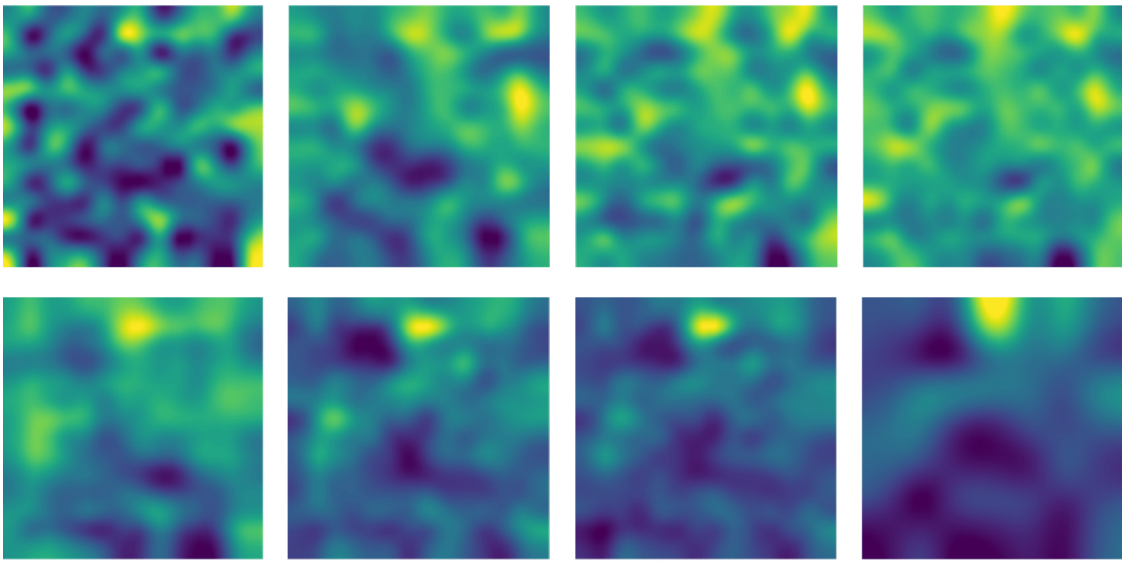

FIGURE 1.10: gradCAM visualization of the last 8 layers for the image at -14.918698, 34.63125076420597.

classes, examining this image in depth may provide useful insights into model decision-making. Figure 1.9 shows both the original input image and the output saliency map produced using backpropagation. Using the saliency map, we can see that both the road and a small settlement are the two human-interpretable features that are highlighted by the model. However, the Grad-CAM output in Figure 1.10 shows a slightly different story. Based on the output heat map for the final four layers of the Vgg-16, the highlighted region appears to be the greener area in the northern region of the model, and to a lesser extent, the road running through the image.

Deciphering the outputs from saliency maps and Grad-CAM is challenging and only provides researchers with suggestions as to what important features are on the ground. Coming up with storylines that policymakers can use to explain model outputs is especially difficult when we compare the results of the classification between the images in Figure 1.9 and similar images like those in Figure 1.11. Each of these images is relatively similar, each coming from the same region in Malawi and containing the same human-interpretable features including roads, areas of greenery, and human settlements. Figure 1.12 shows that many of the same human-interpretable features are highlighted in these cases as well.

FIGURE 1.11: Three similar images from nearby regions in Malawi which were correctly classified

Why then is the image in Figure 1.9 classified incorrectly, while the counter cases in Figure 1.11 are correct? It might be that the varying elevation that we can distinguish in the image was enough to throw off the model, but even our outputs from propagation and Grad-CAM are not enough to provide us with

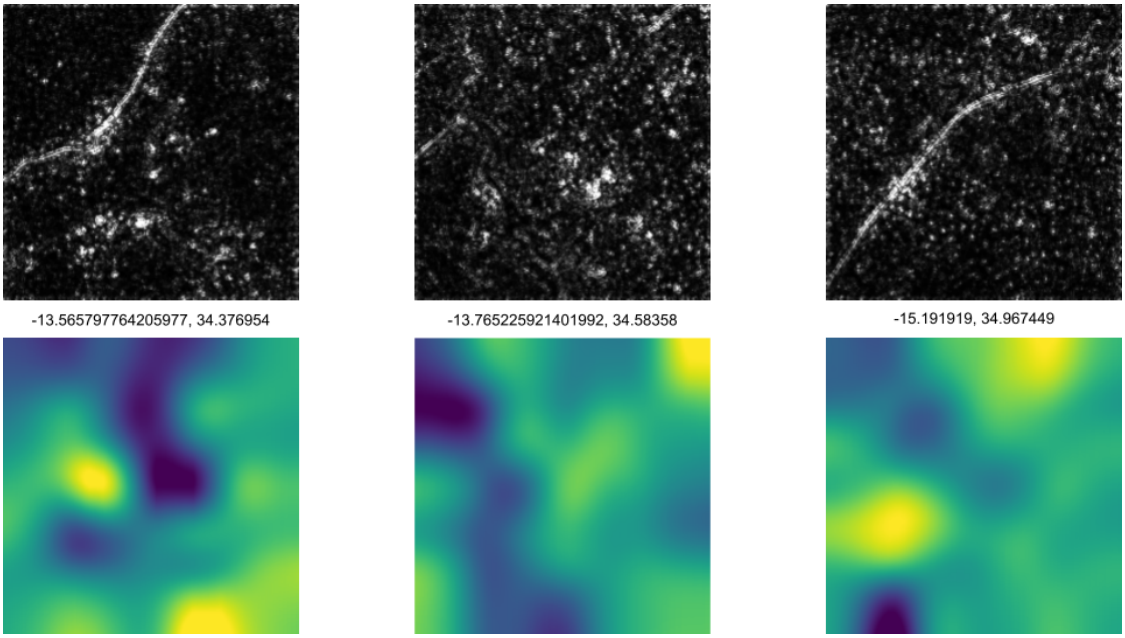

FIGURE 1.12: Output explanations from backpropagation and Grad-CAM activation map of the 28th layer of the previous images.

definitive explanations. This case is an example of the model failing in a case where a human expert may succeed. Human experts with knowledge in the domain may recognize the lightness of the road as an indication that the road is unpaved and the relatively low level of human development as an indicator of the area being of low socioeconomic standing.

6.2 Classification Failures in Urban Areas

Our analysis also showed that the model failed to correctly classify several images in urban areas, predicting a higher level of per-capita asset consumption than the ground truth in our observed cases. Figure 1.13 shows four images collected from the Nigerian cities of Port Harcourt and Onishta which were misclassified by the model as being wealthier than the data indicated. Based on the output saliency maps, we have several hypotheses as to why these misclassifications occurred.

The first two images in Figure 1.13 were collected from Port Harcourt and the saliency map outputs for each clearly show that the edges of Bonny River are

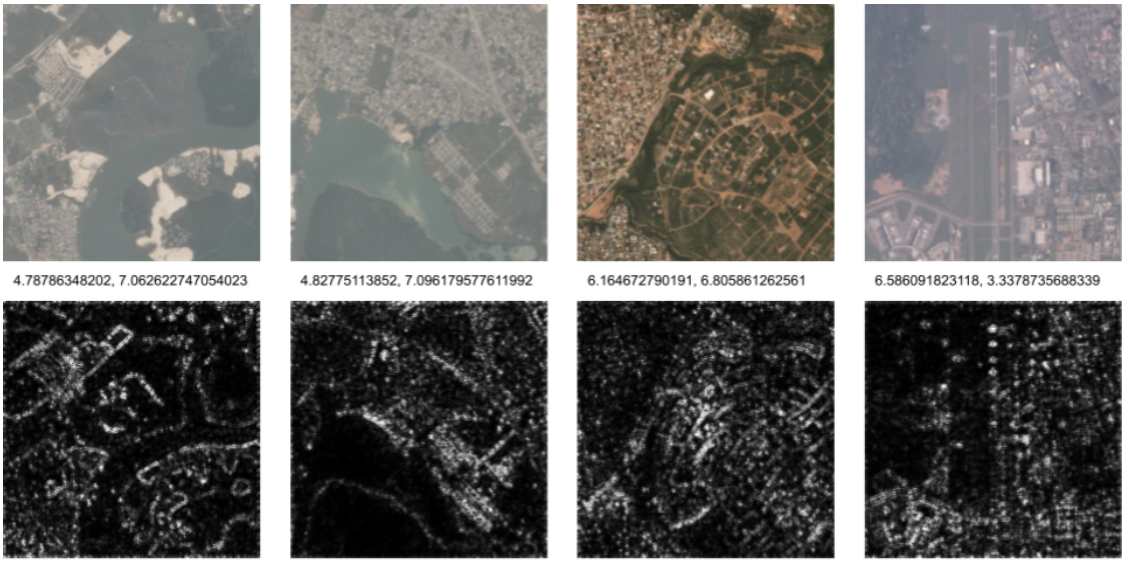

FIGURE 1.13: Urban images misclassified in the Nigerian cities of Port Harcourt and Onitsha

emphasized by the model. In both images, emphasis can be seen on the built-up environment in both industrial and urban areas. The regression model indicated that these features were a significant positive contributor to the classification of the image.

The third and fourth images in Figure 1.13 were collected from Onishta and the output saliency maps highlight the neighborhood streets and buildings in both images. It is interesting to note that roads are highlighted in all four urban images regardless of them being paved or dirt. Each of the saliency maps also highlights the urban buildings clustered in the images.

These cases of urban misclassification are examples of instances in which the model fails in the same way that humans are likely to. It seems likely that the model learns the correlation between a high density of linear features like roads and buildings and higher socioeconomic status. But, because the model can't see under the roofs of these urban neighborhoods and the model is not augmented with nighttime lights or other socioeconomic data, it fails to classify urban slums as being impoverished, as these areas are still largely built up. Thus, impoverished populations that live in densely populated urban areas - especially if housing is interspersed with other structures - are likely to be invisible populations to satellite models.

6.3 Comparing the Performance of Deep Learning Models and Human Experts

In a recent study of domain experts' ability to sort satellite imagery into the appropriate socioeconomic quintile according to DHS survey data on household assets, Ibrahim and Hall, 2022 showed that 19% of images received the same rating from human experts and the DHS survey. Ibrahim and Hall, 2022 write, "tellingly, all the clusters rated as 'Richest' by the domain experts were also rated as 'Richest' in the DHS wealth ranking quintiles. Similarly noteworthy is the fact that none of the 'Poorest' clusters, according to the independent ratings by our domain experts could be found in the 'Richest' quintile of the DHS dataset. This suggests that despite the fact that domain experts do not see below the roof where most household assets can be found, they are, by and large, able to accurately estimate poverty levels merely by examining neighborhood characteristics."

While deep learning models are more accurate than domain experts in classifying satellite imagery based on socioeconomic status, deep learning models fail in unique ways which may lead to portions of the population being misclassified as wealthy, and thereby being excluded from the aid that they need. While none of the images classified by human experts were predicted to be in the furthest quintile from the true quintile, there did exist one image that we analyzed which the model misclassified in this way. This suggests that while domain experts are able to take into account neighborhood characteristics, there may be certain anomalous cases in which our deep learning models fail to do so. We also suggest that in cases of urban poverty, the characteristics which define such poverty may not be visible in satellite imagery.

7 Conclusion

The central question we aimed to answer through our research was *Do there exist implicit biases in our deep learning poverty prediction models which might inhibit certain subjects of the population from receiving poverty relief aid?* The purpose of this paper was to evaluate cases where our model fails and leverage current

interoperability techniques to give us insight into the human-interpretable features highlighted by the model. We trained a Convolutional Neural Network to classify images based on per-capita consumption. We were able to obtain a validation accuracy of 0.74 which is consistent with the current literature. We then evaluated 150 images using backpropagation (Simonyan, Vedaldi, and Zisserman, 2014) and Grad-CAM (Selvaraju et al., 2017) in order to determine which human-interpretable features were emphasized in the model. Our results illustrated that coastlines, buildings, and winding roads were the three features most significant to model classification.

The evaluation of satellite imagery using interpretability techniques demonstrated two things. First, there likely exists some anomalous examples in which the model fails in cases where human experts likely would not. These anomalous cases would benefit from leveraging further interpretability methods and require additional study to determine their prevalence. Additionally, we showed that especially in urban areas, the model fails in the same cases in which human experts are likely to. These findings have implications for urban populations. Our results indicate that poorer populations living in urban areas are at a much higher risk of being misclassified due to neighborhood effects than poor populations situated in rural environments.

In summary, the analysis presented here suggests that there may be implicit biases in CNNs that leverage satellite imagery to predict poverty in data-sparse regions. Specifically, we note three types of bias: (1) a tendency to misclassify urban poor as wealthier; (2) a tendency to classify populations proximate to rivers as "poor"; and (3) a tendency to classify images in which winding roads appear as "wealthy". We recommend further research to explore the sources of these biases and to develop strategies to mitigate them. Future research should focus on improving model generalization and developing more diverse training datasets to reduce the potential for biases. Additionally, we recommend that researchers continue to leverage a variety of interpretability techniques to analyze model bias in poverty prediction and other applications of CNNs.

8 Acknowledgements

We would like to thank the committee Dan Runfola (chair), Matthew Haug, and Jaime Settle. We acknowledge William & Mary Research Computing for providing computational resources, and Matt Kenedy for providing technical support that has contributed to the results reported within this paper. We would also like to thank the faculty and students of the William & Mary geoLab (geolab.wm.edu) for their feedback and support.

Chapter 2

Appendix

Full spreadsheet containing qualitative analysis can be found at: O'Brien, 2023

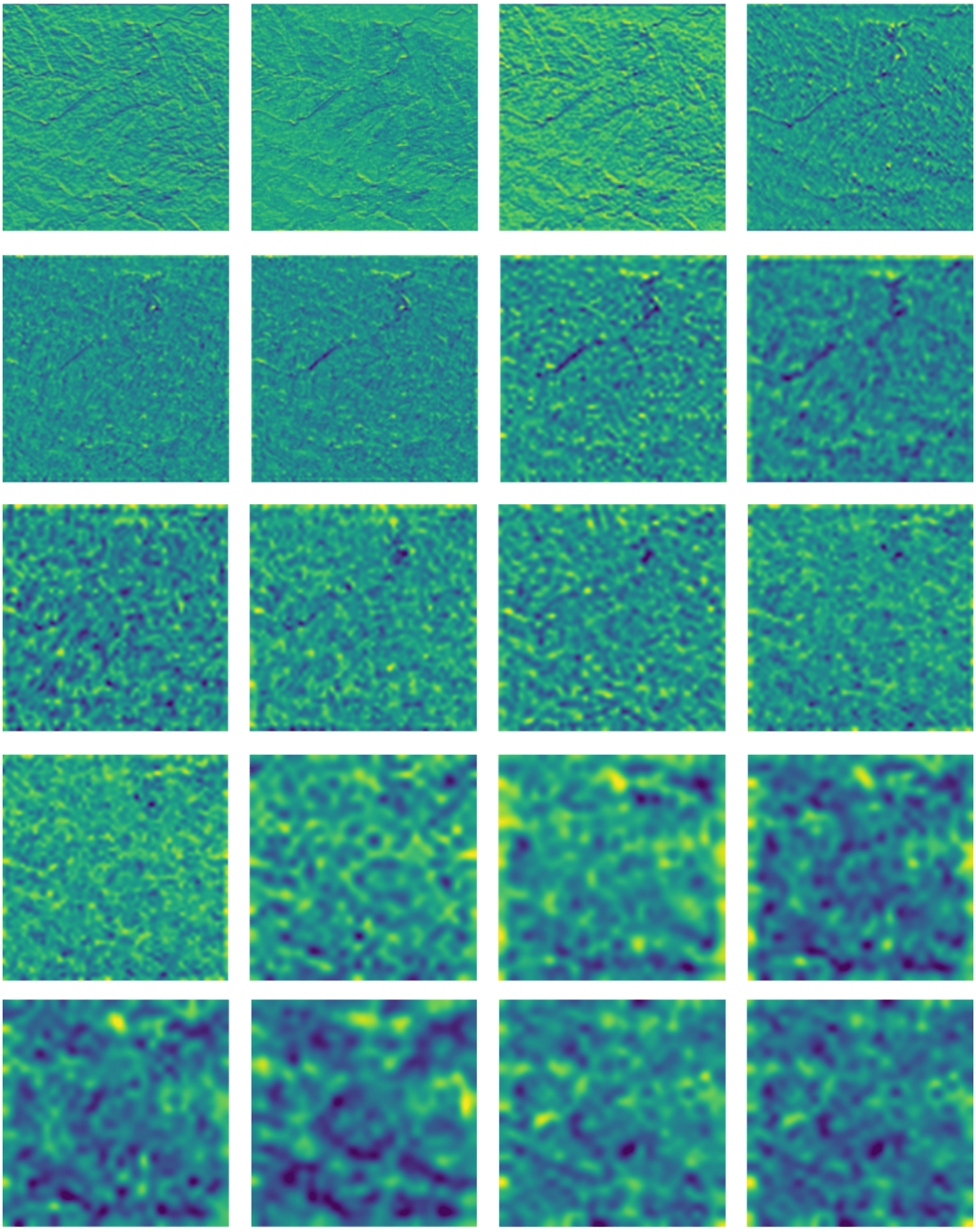

FIGURE 2.1: gradCAM visualization of the first 20 layers for the image at -14.919, 34.631.

Bibliography

Ayush, Kumar et al. (2020). "Generating interpretable poverty maps using object detection in satellite images". In: *arXiv preprint arXiv:2002.01612*.

Ayush, Kumar et al. (2021). "Efficient poverty mapping from high resolution remote sensing images". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 1, pp. 12–20.

Babenko, Boris et al. (2017). "Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico". In: *arXiv preprint arXiv:1711.06323*.

Burke, Marshall et al. (2021). "Using satellite imagery to understand and promote sustainable development". In: *Science* 371.6535, eabe8628.

Castro, Clinton (2019). "What's wrong with machine bias". In: *Ergo, an Open Access Journal of Philosophy* 6.

Castro, Diego A and Mauricio A Álvarez (2023). "Predicting socioeconomic indicators using transfer learning on imagery data: an application in Brazil". In: *Geojournal* 88.1, pp. 1081–1102.

Chakraborty, Supriyo et al. (2017). "Interpretability of deep learning models: A survey of results". In: *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBD* IEEE, pp. 1–6.

Chen, Derek (2017). *Temporal poverty prediction using satellite imagery*.

Chi, Guanghua et al. (2022). "Microestimates of wealth for all low-and middle-income countries". In: *Proceedings of the National Academy of Sciences* 119.3, e2113658119.

Crawford, Michael et al. (2015). "Survey of review spam detection using machine learning techniques". In: *Journal of Big Data* 2.1, pp. 1–24.

Daoud, Adel et al. (2021). "Using satellites and artificial intelligence to measure health and material-living standards in India". In: *arXiv preprint arXiv:2202.00109*.

Deviyani, Athiya (2022). "Assessing Dataset Bias in Computer Vision". In: *arXiv preprint arXiv:2205.01811*.

Espın-Noboa, Lisette, János Kertész, and Márton Karsai (2022). "Challenges of inferring high-resolution poverty maps with multimodal data". In.

Ferguson, Andrew Guthrie (2017). "Rise of Big Data Policing, The". In: *Rise of Big Data Policing, The*. New York University Press.

Garvie, Clare and Jonathan Frankle (2016). "Facial-recognition software might have a racial bias problem". In: *The Atlantic* 7.

Gilpin, Leilani H et al. (2022). ""' Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI". In: *arXiv preprint arXiv:2207.00007*.

Hall, Ola, Mattias Ohlsson, and Thorsteinn Rögnvaldsson (2022). "A review of explainable AI in the satellite data, deep machine learning, and human poverty domain". In: *Patterns* 3.10, p. 100600.

Hasan, Reem Ibrahim, Suhaila Mohd Yusuf, and Laith Alzubaidi (2020). "Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion". In: *Plants* 9.10, p. 1302.

Head, Andrew et al. (2017). "Can human development be measured with satellite imagery?" In: *Ictd* 17, pp. 16–19.

Hofer, Martin et al. (2020). "Applying Artificial Intelligence on Satellite Imagery to Compile Granular Poverty Statistics". In: *Asian Development Bank Economics Working Paper Series* 629.

Huang, Luna Yue, Solomon M Hsiang, and Marco Gonzalez-Navarro (2021). *Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs*. Tech. rep. National Bureau of Economic Research.

Ibrahim, Wahab and Ola Hall (2022). "Welfare estimations from imagery. A test of domain experts ability to rate poverty from visual inspection of satellite imagery". In: *arXiv preprint arXiv:2210.08785*.

Irvin, Jeremy, Dillon Laird, and Pranav Rajpurkar (2017). *Using satellite imagery to predict health*. Tech. rep. Technical report, Stanford University, Department of Computer Science.

Jarry, Robin et al. (2021). "Assessment of CNN-based methods for poverty estimation from satellite images". In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*. Springer, pp. 550–565.

Jean, Neal et al. (2016). "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301, pp. 790–794.

Jean, Neal et al. (2019). "Tile2vec: Unsupervised representation learning for spatially distributed data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 3967–3974.

Ker, Justin et al. (2017). "Deep learning applications in medical image analysis". In: *Ieee Access* 6, pp. 9375–9389.

Khosla, Aditya et al. (2012). "Undoing the damage of dataset bias". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*. Springer, pp. 158–171.

Kim, Jae Hyun et al. (2016). *Incorporating spatial context and fine-grained detail from satellite imagery to predict poverty*. Tech. rep. Working paper, Stanford University.

Kondmann, Lukas and Xiao Xiang Zhu (2020). "Measuring changes in poverty with deep learning and satellite imagery". In.

Krishna, Satyapriya et al. (2022). "The disagreement problem in explainable machine learning: A practitioner's perspective". In: *arXiv preprint arXiv:2202.01602*.

Lee, Kamwoo and Jeanine Braithwaite (2022). "High-resolution poverty maps in sub-saharan africa". In: *World Development* 159, p. 106028.

Lipton, Zachary C (2018). "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Liu, Haoyu et al. (2021). "Nightlight as a proxy of economic indicators: Fine-grained gdp inference around chinese mainland via attention-augmented cnn from daytime satellite imagery". In: *Remote Sensing* 13.11, p. 2067.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Mathur, Jatin (2020). *predicting-poverty-replication*. URL: http://github.com/jmather625/predicting-poverty-replication (visited on 08/30/2022).

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267, pp. 1–38.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Ni, Ye et al. (2020). "An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction". In: *IEEE Geoscience and Remote Sensing Letters* 18.9, pp. 1545–1549.

O'Brien, Joseph (2023). *Honors23*. URL: http://github.com/joeeobrienn/Honors23 (visited on 04/25/2023).

Pandey, Shailesh, Tushar Agarwal, and Narayanan C Krishnan (2018). "Multi-task deep learning for predicting poverty from satellite images". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Perez, Anthony et al. (2017). "Poverty prediction with public landsat 7 satellite imagery and machine learning". In: *arXiv preprint arXiv:1711.03654*.

Perez, Anthony et al. (2019). "Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty". In: *arXiv preprint arXiv:1902.11110*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""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*, pp. 1135–1144.

Robinson, Caleb, Fred Hohman, and Bistra Dilkina (2017). "A deep learning approach for population estimation from satellite imagery". In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pp. 47–54.

Runfola, D, A Stefanidis, and H Baier (2022). "Using satellite data and deep learning to estimate educational outcomes in data-sparse environments". In: *Remote Sensing Letters* 13.1, pp. 87–97.

Runfola, Daniel et al. (2022). "Deep learning fusion of satellite and social information to estimate human migratory flows". In: *Transactions in GIS* 26.6, pp. 2495–2518.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034*.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke (2016). "Deep learning for mortgage risk". In: *arXiv preprint arXiv:1607.02470*.

Tammina, Srikanth (2019). "Transfer learning using vgg-16 with deep convolutional neural network for classifying images". In: *International Journal of Scientific and Research Publications (IJSRP)* 9.10, pp. 143–150.

Tan, Yumin et al. (2020). "Combining residual neural networks and feature pyramid networks to estimate poverty using multisource remote sensing data". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13, pp. 553–565.

Tang, Binh, Yanyan Liu, and David S Matteson (2022). "Predicting poverty with vegetation index". In: *Applied Economic Perspectives and Policy* 44.2, pp. 930–945.

Terhörst, Philipp et al. (2021). "A comprehensive study on face recognition biases beyond demographics". In: *IEEE Transactions on Technology and Society* 3.1, pp. 16–30.

Tingzon, Isabelle et al. (2019). "MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION." In: *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Tommasi, Tatiana et al. (2017). "A deeper look at dataset bias". In: *Domain adaptation in computer vision applications*, pp. 37–55.

Vardi, Gal (2022). "On the implicit bias in deep-learning algorithms". In: *arXiv preprint arXiv:2208.12591*.

Wu, Peng and Yumin Tan (2019a). "Estimation of economic indicators using residual neural network ResNet50". In: *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 206–209.

— (2019b). "Estimation of poverty based on remote sensing image and convolutional neural network". In: *Advances in Remote Sensing* 8.4, pp. 89–98.

Xie, Michael et al. (2016). "Transfer learning from deep features for remote sensing and poverty mapping". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1.

Yeh, Christopher et al. (2020). "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". In: *Nature communications* 11.1, p. 2583.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, pp. 818–833.

Zhang, Quanshi, Wenguan Wang, and Song-Chun Zhu (2018). "Examining CNN representations with respect to dataset bias". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Zhao, Xizhi et al. (2019). "Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh". In: *Remote Sensing* 11.4, p. 375.

Zhou, Bolei et al. (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.