Patterns

CellPress
OPEN ACCESS

Review

A review of explainable AI in the satellite data, deep machine learning, and human poverty domain

Ola Hall,[1,*] Mattias Ohlsson,[2,3] and Thorsteinn Rögnvaldsson[2]
[1]Department of Human Geography, Lund University, Lund, Sweden
[2]Center for Applied Intelligent Systems Research, Halmstad University, Halmstad, Sweden
[3]Division of Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden
*Correspondence: ola.hall@keg.lu.se
https://doi.org/10.1016/j.patter.2022.100600

THE BIGGER PICTURE Research has shown that certain aspects of human poverty and welfare can be measured with deep machine learning in combination with satellite imagery. High hopes have been expressed from policy and decision makers, but downstream applications are still rare. We suggest that one obstacle to overcome is the lack of explainability in some parts of the scientific process. Future directions would be to develop methods to understand which features in an image that triggers a certain response and relate that to domain knowledge; is the response and features in accordance with theory within the field(s)? Suggested reading: Roscher, Ribana, et al. "Explainable machine learning for scientific insights and discoveries." *IEEE Access* 8 (2020):42;200–42216.

SUMMARY

Recent advances in artificial intelligence and deep machine learning have created a step change in how to measure human development indicators, in particular asset-based poverty. The combination of satellite imagery and deep machine learning now has the capability to estimate some types of poverty at a level close to what is achieved with traditional household surveys. An increasingly important issue beyond static estimations is whether this technology can contribute to scientific discovery and, consequently, new knowledge in the poverty and welfare domain. A foundation for achieving scientific insights is domain knowledge, which in turn translates into explainability and scientific consistency. We perform an integrative literature review focusing on three core elements relevant in this context—transparency, interpretability, and explainability—and investigate how they relate to the poverty, machine learning, and satellite imagery nexus. Our inclusion criteria for papers are that they cover poverty/wealth prediction, using survey data as the basis for the ground truth poverty/wealth estimates, be applicable to both urban and rural settings, use satellite images as the basis for at least some of the inputs (features), and the method should include deep neural networks. Our review of 32 papers shows that the status of the three core elements of explainable machine learning (transparency, interpretability, and domain knowledge) is varied and does not completely fulfill the requirements set up for scientific insights and discoveries. We argue that explainability is essential to support wider dissemination and acceptance of this research in the development community and that explainability means more than just interpretability.

INTRODUCTION

Weather events, such as the current drought in East Africa leaving more than 13 million people severely food and water insecure, the COVID-19 pandemic, and food price shocks due to war and local conflicts, affect already vulnerable and disadvantaged communities. Identifying those with the greatest needs is a challenging task aggravated by significant information gaps typical for many poor regions of the world. Lack of accurate and repeated measurements has long hampered effective relief efforts and progress toward understanding the determinants of observed outcomes and ultimately progress toward sustainable development.[1]

Programs tailored for measuring different aspects of living standards at a household level have long been the primary tool to identify people at risk and long-term development.[2] Since 1984 the Demographic and Health Surveys (DHS) program has, together with around 90 developing countries, collected national representative data on fertility, reproductive health, maternal health, child health, immunization and survival, and welfare in general (https://dhsprogram.com/). The objective of the DHS program is to improve and formalize surveys and the use of data together with individual countries for program monitoring and evaluation and policy development decisions. Similar, but not identical, the Living Standards Measurement Study (LSMS) is a project that was initiated in 1980 (https://www.worldbank.org/en/programs/lsms). Although DHS and LSMS share the goal of monitoring, evaluation, and policy development decisions, they aim to move beyond simple measurement rates and toward an explanation and increased understanding of outcomes. Together they have served as workhorses for understanding social and economic progress for more than four decades.

While household-level surveys are the key tool for information on human welfare and outcomes, they are also increasingly criticized. They are expensive and laborious, with a typical LSMS or DHS survey round for one country costing close to 2 million USD and population census being 10 or 100 times more expensive.[3,4] Related, many countries conduct surveys infrequently, sometimes due to the associated high costs and an unwilling regime. Another implication is that surveys of this kind are generally only representative of nations and sometimes regions, making it difficult to access information in sub-regional locations.

The traditional tools for gathering data on human development are now supplemented with digital data sources and tools. Central to what the UN has named the "data revolution" has been the growing abundance and quality of satellite data, together with developments in machine learning (ML) over the last two decades, in particular deep learning and convolutional neural networks (CNNs) for image analysis. This development has been possible because of considerable increases in labeled datasets, massive improvements in computer hardware, and substantial developments in ML algorithms that can exploit the hardware. Today, the ImageNet database has more than 14 million images (https://www.image-net.org/) that can be used to train deep neural networks and the CIFAR-10 database, a subset of the 80 million tiny images dataset, has more than 60,000 labeled images (https://www.cs.toronto.edu/ kriz/cifar.html). Ten years ago, the best non-human results in the ImageNet competition were about 25% wrong classifications in the "top 5" category;[5] today, they are in the single percentages and much better than human performance.[6] The same development has been seen in the CIFAR-10 competition; the deep learning ML approaches surpassed human performance roughly 5 years ago.[7,8] Training state-of-the-art deep networks on the ImageNet dataset took days or weeks a decade ago but takes minutes or hours today.[9]

An interesting discovery with deep CNNs is that they learn high-level features in their top layers that can be transferred to a related domain with fewer data, and they provide improved performance compared with just using the data from that domain.[10] Such transfer learning was demonstrated on deep CNNs trained on the ImageNet data and then tuned to satellite image data in a seminal paper by Jean et al. in 2016.[11] This inspired many following studies and, for some human outcome indicators, e.g., asset-based poverty, ML approaches combined with satellite imagery are now close to matching the performance of survey data.

The most recent review of the field is by Burke et al. (2021).[2] They performed a broad review of the rapidly growing literature regarding satellite imagery and measurements of different human outcomes, with specific attention to those who apply artificial intelligence to images. Their work focused on four domains where satellite-based measurements have been particularly successful: smallholder agriculture, population estimation, informal settlements, and economic livelihoods. The four domains are naturally intertwined in terms of focusing on humans and human outcomes but, in terms of research, they form rather distinct categories. Remote sensing in agriculture is a vast research field that involves a range of methods and applications adapted to varying contexts, such as precision agriculture in developed regions of the world and crop yield studies in poor regions. Smallholder agriculture, typical for large parts of Africa, is specifically challenging for remote sensing due to small field sizes, complex environments, and scarcity of reference data. There are several reviews on remote sensing in agriculture, for example, Atzberger[12] and Weiss et al.[13] Knowing where people are located is vital information and a starting point for many applications. For a review of five satellite-based population datasets, see Bustos et al.[14] The detection of informal settlements or slums is a field that has received attention recently and can be directly linked to the well-being of people. For a review, see Kuffer et al.[15]

The domain for our concerns is the one Burke and co-workers categorize as economic livelihoods. Studies in this category aim to predict local-level human outcomes of different types and over varying geographies. Most studies predict wealth with a focus on the developing world (especially Sub-Saharan Africa) [SSA]) and are benchmarked against DHS data. The work in this category originates from some key publications, Xie et al.[16] and Jean et al.[11] In an attempt to quantify the accumulated performance, Burke et al.[2] found that often more than 75% of the variation in the survey measured performance could be explained, a number that seems to increase over time. The overall conclusion from Burke et al. is that AI-based methods outperform earlier methods whenever they are deployed. The success has spurred high hopes in both research and policy communities, and there is a growing interest in how these findings can be put to work.[2,17]

This paper investigates the potential to learn something new from using deep ML and satellite imagery in the specific domain of poverty and welfare. The aim is linked to recent debates on explainability in computer sciences where "there is a recent and ongoing high demand in the ML-community for understanding the way a specific model operates and the underlying reasons for the decisions made by the model" (Roscher et al.[18]). Explainability is to be understood as a prerequisite to ensure the scientific value, and thereby deepening our understanding and ultimately providing new scientific discoveries. Therefore, we use the framework from Roscher et al. to structure and review the concepts of transparency, interpretability, explainability, and domain knowledge (see Tables 1, 2, and 3).

A pertinent question is, then, can these new tools be used to provide scientific insight on a large scale concerning why people are poor? Can they provide more basis for knowledge than "just" an estimate of poverty? The geography of poverty is reasonably well known; *why* people are poor is a more complicated question

Table 1. Descriptions of how model, design, and algorithmic transparency were evaluated for the reviewed papers

Model transparency	Design transparency	Algorithmic transparency
The model is mathematically transparent (e.g., functions, size, deterministic …)	Decisions in the design (e.g., kernels, layers, units, …) are clearly described and motivated.	The way to find the solution is "unique," the solution can be found again (e.g., stopping criteria …)
The paper should describe the model structure so an expert can write the model down mathematically. If this is the case, then it is valued as "mathematically transparent" (green). If the paper mentions the use of a method but nothing about what the final model looks like, then it is valued as "not mathematically transparent" (red). If the paper is somewhere between those two cases, e.g., that the feature extraction would be hard to reproduce, but the remaining model is ok, then it is valued as "somewhat mathematically transparent" (yellow).	The paper should describe the model design choices well enough so that an expert can understand and repeat those choices. If this is the case, then the model is valued as "design transparent" (green). If no motivation or specification is given on the design choices, then the model is valued as "not design transparent" (red). If the description in the paper is in between those two, some choices are described well, others not so well, then the paper is valued as "somewhat design transparent" (yellow).	In ML there is an acceptable level of non-uniqueness in the solution. Parameters may not be identical, but the functional result of repeated training can be almost identical. If the algorithm and training method used in the paper are described such that repeated runs should produce a functionally almost identical result, then the paper is labeled "algorithmically transparent" (green). If important issues, like stopping criteria, are not described, then the paper is valued as "not algorithmically transparent" (red). Anything in between is valued as "somewhat algorithmically transparent" (yellow).

to answer. We investigate the status and potential of these questions by reviewing all the papers from the beginning, starting almost a decade ago. Common for this body of work is that targets, training, and evaluation are all conducted in relation to household survey data, particularly DHS or LSMS. We do not include other research fields where ML applications have succeeded, such as slum detection or poverty mapping in general.

Remote sensing and ML predictions of poverty

In the mid-1990s, the National Aeronautics and Space Administration approached the research community in an effort to realize the potential of satellite imagery—specifically addressing the social sciences. High hopes were expressed in "People and pixels: linking remote sensing and social science."[19] However, the results have been meager, and their added value questioned[20,21] until recently. The most influential work where satellite data are applied to social and economic research originates from the Defense Meteorological Satellite Program Operational Line-Scan System after scholars, already in the mid-1970s, observed that imagery showed the extent and intensity of human settlements. Data with this capability are usually referred to as night time lights (NTLs). Technical limitations with data storage and processing power hampered the accessibility of imagery and development until the 1990s. NTL data have been publicly available from 1992 and onward.

Early work observed that regions emitting high levels of NTL were also associated with high economic output.[22] Henderson and others used data on annual global NTLs and showed the linkage between artificial light and economic activity.[23–26] While the relation was observed much earlier by Elvidge et al.[22] the previous authors who detailed the econometrics, an ongoing work[27] showed that such data correlate closely with detailed records of wage income in Sweden. NTL data has been applied and evaluated in some of the world's poorest regions. The work[26] showed that NTL correlated with asset-based measures of wealth for 37 countries in Africa, but at the same time underlined the observation that NTLs underperform in the poorest regions, which was confirmed in a study in Burkina Faso.[28] This is usually attributed to the low light levels found in agricultural-based economies and poor regions, which cannot be separated from noise in the data.[24] It has also been observed that NTLs have difficulties distinguishing between poor and densely populated areas; and wealthy sparsely populated areas.[11] Improvements are observed with the spatially and radiometrically improved sensor Visible Infrared Imaging Radiometer Suite.[29]

These limitations inspired recent papers that use daytime satellite imagery to measure poverty in developing countries, particularly in SSA. Pioneering works by Xie et al.[16] and Jean et al.[11] combined the power of NTL with daytime satellite imagery and recent tools in ML. To circumvent the lack of labeled training data, they applied a two-step transfer learning approach to five countries in SSA using NTL intensity levels as labels. They improved R2 by more than 10% compared with using NTL alone. This approach has been elaborated further in several papers and for countries outside Africa, for example, Sri Lanka, China, and India, and with steadily improving performance metrics.[2] The benchmark indicator for these studies is the wealth index (WI) from the DHS. The index is a principal-component analysis of items easily observable from the surveying officer's perspective, such as access to water, phones, and bicycles. On the other hand, Head et al.[30] have shown that this method does not generalize in the same way that other measures of development predict access to drinking water and various health indicators. Other measures, such as the consumption index from the LSMS are not predictive at the same level as asset-based indices (see below). Overall, satellite image-based poverty predictions can now explain more than half of the variation and sometimes up to 85% of the survey-measured poverty. In this paper, we focus the review on papers that build on the seminal texts from 2015/2016.

Measuring poverty

Poverty measures have been a concern for research and society for more than a century.[31] Some of the studies in our review include more than one indicator, but the common denominator is that they all predict poverty, or economic livelihood to follow

Table 2. Descriptions of how interpretability and algorithmic explainability were evaluated in the reviewed papers

Interpretability	Algorithmic explainability
The properties of the final model are described in a way that is understandable to a human.	What, how, and why?
The paper should make an effort to describe the properties of the model in an understandable way. In the case of a linear model, this is straightforward. For simpler decision trees, this is also straightforward. For more complicated models, saliency maps or heat maps can be used to illustrate what the model reacts to. If such a description is well explained in the paper, it is valued as an "interpretable" model (green). If there is no attempt to describe the models' properties, then it is valued as "uninterpretable" (red). Cases between these extremes are valued as "partly interpretable" (yellow).	Does the paper make an attempt to explain why a certain prediction is made? For example, why are some villages considered poor and others not? What would need to change in the satellite data for a village to move from poor to not so poor? If the paper makes an attempt at this, or if the answer is obvious from the model structure (e.g., a linear model), then it is valued as "explainable" (green). If there is no discussion at all in the paper on this and the model is not straightforward to explain, then it is valued as "not explainable" (red). Cases between these two are valued as "partly explainable" (yellow).

the Burke et al. categorization,[2] and which is also the overarching aim of most of the studies reviewed.

Poverty can be classified in several ways. One way is to divide poverty into absolute and relative terms. The former uses poverty lines with constant real value as in the World Bank definition of extreme poverty (i.e., those who live on less than $1.90 a day). The latter uses relative measures for which the poverty line varies as a function of a set proportion concerning the current mean (or median).[31] Examples of both can be found in the reviewed literature. Further division concerns persistent and transient poverty, place-based and individual poverty, and urban and rural poverty.[32] Place-based poverty tends to be persistent and, accordingly, individual poverty transient.

There is a fundamental difference between stock measures (assets) and flow measures (income, consumption, and expenditures). While it is desirable to have data on household income and expenditures, they are limited by factors, such as seasonality, misreporting, and volatility.[33] Household assets are easier to collect as they involve items that are easily observed by the surveying officer and are closely linked to long-term welfare status. A household's assets are more obvious targets for remote sensing-based methods than most flow measures. The majority of studies we have reviewed target asset wealth as an indicator of poverty. Several studies attempt to estimate consumption expenditure but with less success. There are also attempts to estimate the global multidimensional poverty index and also poverty rates derived from censuses (see Table 4).

The WI found in most of the literature is the DHS WI, a composite measure of a household's cumulative living standard. The WI is calculated based on a household's ownership of selected assets, such as televisions, bicycles, and materials used for housing construction (flooring and tiling), as well as types of water access and sanitation facilities. Some assets are clearly hidden from space-borne sensors. The WI places individual households on a continuous scale of relative wealth, making comparisons difficult between countries. DHS separates all interviewed households into five wealth quintiles to compare the influence of wealth on various population, health, and nutrition indicators. Later versions also include land holdings and farm animals. To protect the integrity of respondents, the DHS village cluster coordinates are randomly displaced up to 10 km. A similar approach is used for LSMS.[45]

EXPLAINABLE AI

Recently, there have been several surveys on explainable AI (XAI) methods and terminology. Most relevant for our discussion is the one by Roscher et al.[18] where they discuss requirements for using ML for scientific discovery and organize them into three core elements.

They make a useful distinction between *transparency*, *interpretability*, and *explainability*, where transparency considers the ML approach, interpretability considers the ML model together with data, and explainability considers the model, the data, and human involvement.

Transparency

In general, ML models are transparent to the extent that they are mathematically well defined; equations can be written down that describe what they do. However, ML models tend not to be transparent in the sense that it is easy to understand why certain model design choices were made (e.g., number of layers, activation functions, regularization, training algorithm). Transparency relates to the processes for constructing the ML model, e.g., the final model itself, methods for model structure choices, and for fitting the parameters. If these can all be well described and motivated then the ML model is transparent. Following Lipton[63] and Roscher et al.[18] it is necessary to divide transparency into three parts: model transparency, design transparency, and algorithmic transparency.

Model transparency relates to the transparency of the structure of the model (e.g., the number of layers, activation functions, kernel functions, number of decision trees in a random forest, splitting criteria). Design transparency refers to design choices made when constructing the ML algorithm: are those choices understandable, well motivated, and replicable? Examples include selecting neural network architecture, activation functions, training time, batch sizes, training algorithm, which may all affect the final ML model. Algorithmic transparency relates to the uniqueness of the final solution. Is the result reproducible even if all design choices are reported thoroughly? Frequently, there are several local minima where a model can get stuck and the result of two training sessions can end up quite different. If that is the case, then the algorithm is not transparent.

Table 3. Descriptions of how domain knowledge was used regarding data, hypothesis, training, and the final model were evaluated in the reviewed papers

Data (features)	Hypothesis (model structure)	Training (loss function)	Final model (constraints)
Does the paper employ domain knowledge in selecting or engineering features? If this is the case, then it is valued as "domain knowledge in data" (green). If the work relies completely on learning from data without any prior knowledge, then it is valued as "no domain knowledge in data" (red). Cases in between, e.g., using both domain knowledge features and pure learning from data, are labeled as "some domain knowledge in data" (yellow). For example, fine-tuning a CNN with night time light data are labeled as "some domain knowledge in data."	Is there some hypothesis built into the model? For example, do certain symmetries or some features positively or negatively impact the prediction? If this is the case, then the work is valued as a "hypothesis included in model" (green). If there is no such hypothesis, then it is valued as "no hypothesis included in model" (red). In unclear cases, or if there is some weak hypothesis used, then it is valued as "some hypothesis included in model" (yellow).	Does the training procedure utilize domain knowledge? For example, that the loss function should have a specific form for this problem. If this is the case, then the work is valued as "domain knowledge in training" (green). If a standard loss function and no domain knowledge are used in the training, then it is valued as "no domain knowledge in training" (red). In unclear cases, then it is valued as "some domain knowledge in the training" (yellow).	Is it checked if the final model fulfills known or expected relationships or constraints? The simplest example can be that development index values should be positive. Others could be that strong increases in predicted values should be manifested by certain changes in the features. If there is a proper discussion on this in the paper, then it is valued as "domain knowledge checked in final model output" (green). If there is no such discussion in the paper, then it is valued as "no domain knowledge checked in final model output" (red). In cases between these two, it is valued as "some domain knowledge checked in final model output" (yellow).

Interpretability

Interpretability is often what is meant when explaining a model. Interpretability concerns describing the properties of an ML model to a person, i.e., "The mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of."[64] Methods for interpretability try to determine and show which input data (or part of the input data) was responsible for the model prediction. Some models are intrinsically interpretable, e.g., linear models, and can be preferred even though they may be less accurate. Methods for interpreting ML models include the Shapley additive explanations (SHAP) technique, and many others. If the output (decision) from an ML model can be described locally, e.g., by heat maps, filter responses, local expansions, then one can say that the model is interpretable. See "interpretability methods for images" below for other methods.

Explainability

Explainability is a collection of interpretations with further contextual information. It deals with causality, the "what," "how," and "why" questions.[65] It may not be enough to look at a single data point and an interpretation of the resulting prediction to explain a model, e.g., which pixels in an image were important for a prediction. Knowledge creation for scientific purposes requires understanding the relationships encoded in the model, using concepts understood by the scientific community, and agreeing with (and using) prior domain knowledge. This often means simplifying; Can a (large) group of features be collected together into one or a few concepts? Can relations be expressed with few and simple operations? And so on. With this definition of explainability, it is clear that much work remains to achieve explainable ML models. Essentially all work so far has focused on interpretability.

Domain knowledge

Domain knowledge refers to the background knowledge of the field or environment to which the methods are applied. Three aspects are involved, according to van Rueden et al.[66]: type of knowledge, representation, and transformation of knowledge and integration of the above into the ML approach. They arrange different types of knowledge along a continuum from sciences to engineering toward individuals' intuition. Domain knowledge can be integrated into an ML approach in the training data, hypothesis, training algorithm, and final model.

Interpretability methods for images

Several review papers exist on interpretable deep learning models for image analysis, such as Van der Velden et al.[67] and Gulum[68] for images from the medical domain.

A common approach to explaining a deep-learning imaging model is to create attribution maps. The attribution values can be interpreted as the contribution or relevance of each input feature for the given task. In the case of images where input features are the actual pixel values, attribution maps are often presented as an image of the same size as the input images and provide a direct visual explanation of the method. Many attribution methods can be applied to trained deep-learning models, such as the family of CNNs, without any modifications to the underlying architectures or learning procedures. This makes them ideal for many applications also outside of the medical domain, such as satellite imaging.

Attribution maps can be created using two broad approaches, perturbation (e.g., occlusion)-based methods or back-propagation (e.g., gradient)-based methods. The former approach is modifying the input image and measuring its effect on the model's output. The perturbation can be modifying individual

Table 4. A list of all reviewed papers including comments on methods and data used

Reference	Year	Method	Data
Chen et al.[34]	2016	First, ResNet50 Convolutional Neural Network (CNN) model trained on ImageNet. The second step is done with linear ridge regression.	Input: first, eight bands from year-averaged landsat-7 satellite images. Second, the features from the CNN. Target: first, night time lights. Second, a normalized wealth score computed from survey data collected by the World Bank in the Living Standard Measurement Study (LSMS).
Jean et al.[11]	2016	The VGG-F CNN pre-trained on ImageNet data, fine-tuned on night time lights. Ridge regression used for the final model from CNN features to poverty/well-being indices.	Input: satellite image data from Google Static Maps API, zoom level 16 (several zoom levels are tried). Target: two are used. Consumption expenditure as measured in the World Bank's LSMS. Expenditures are averaged over clusters and the log expenditure is modeled. Household asset score taken from the Demographic and Health Surveys (DHS).
Kim et al.[35]	2016	ResNet-50 CNN pre-trained on ImageNet and then tuned on night time lights.	Input: satellite image data from Google Static Maps API, zoom level 14, 16, and 18 (all three levels are used). Target: asset wealth index computed from the DHS.
Xie et al.[16]	2016	First, the VGG-F CNN model, pre-trained on ImageNet. Second, logistic regression.	Input: first, satellite images from the Google Static Maps API, at zoom level 16. Second, the features from the CNN. Target: first, night time light intensities from The National Oceanic and Atmospheric Administration (NOAA). Second, data with binary poverty labels from the LSMS survey conducted in Uganda (Uganda Bureau of Statistics 2012).
Babenko et al.[36]	2017	GoogleNet CNN model. Trained directly to predict poverty.	Input: digital globe and planet satellite imagery (RGB). High-res and low-res. Target: survey data from the 2014 MCS-ENIGH. Income per adult equivalent. Three poverty groups (fraction of households living in poverty). Poverty lines.
Head et al.[30]	2017	VGG16 CNN, pre-trained on ImageNet but fine-tuned with night time lights.	Input: satellite images from Google Static Maps. "Low res." NOAA night time light images from the Defense Meteorological Satellite Program Operational Line-Scan System (DMSP-OLS) website are used for fine-tuning. Target: DHS data. Several indices from these.
Irvin et al.[37]	2017	Tries both "standard" CNN and a ResNet CNN network. Tries both training the ResNet from scratch and having one pre-trained on ImageNet data but fine-tuned on their task. Tests with averaging outputs from image tiles, or using more advanced recurrent neural networks for combining (LSTM).	Input: satellite images from Google Maps Static API, zoom level 15. Target: DHS data. Poverty prediction with wealth index, split into four categories. Malnutrition prediction using height and weight for age scores, split into six categories.
Perez et al.[38]	2017	CNN of type ResNet and VGG-Net. Pre-trained on ImageNet data for RBG bands, otherwise trained from scratch. Try both gradient boosted trees and linear ridge regression for predicting Asset Wealth Index (AWI) from the CNN features, and also (as comparison) directly from night time light values.	Input: first, multispectral satellite imagery from Landsat 7 (several spectral bands tried). Second, CNN features. Target: first, predicting class of night time lights (three classes). Second, AWI from DHS data.

(*Continued on next page*)

Table 4. *Continued*

Reference	Year	Method	Data
Pandey et al.[39]	2018	CNN architecture for the first task. Trained from scratch. For the second task a multilayer perceptron.	Input: first, satellite images from Google Static Maps API, zoom level 16. Images only of villages. Second, predicted roof material, source of lighting, and source of drinking water. Target: first, the roof material, source of lighting, and source of drinking water in a region. In a second model, the household income level in a region. The information comes from the 2011 Census of India and Socio-Economic Caste Census of 2011.
Perez et al.[40]	2019	CNN with weighted Generative Adversarial Network (WGAN), for the first step. The second step is done with ridge regression.	Input: first, multispectral (nine bands) imagery from Landsat 7 (Enhanced Thematic Mapper Plus [ETM+]). RBG images are pan-sharpened (to get 15 m resolution). Second, the features from the CNN. Target: it is not described fully what they train the CNN to predict, but it includes night time lights. For a second, linear model, they use the AWI from the DHS data.
Tingzon et al.[41]	2019	One approach using VGG16 CNN, pre-trained on the ImageNet data and fine-tuned on night time lights, ridge regression used on the final features. Another approach using Random Forest (RF) on OpenStreet Map (OSM) engineered features (road type, building type, and points of interest). The results from the two approaches are compared.	Input: daytime satellite images from Google Static Maps, zoom level 17. OSM data from GeoFabrik, with OSM features then engineered by the group themselves. Target: CNN approach first uses night time luminosity data from Visible Infrared Imaging Radiometer Suite Day/Night Band (VIIRS-DNB), categorized into five classes. Then wealth index (and other indices) from the Philippine DHS from 2017. The ONS approach builds an RF model directly for the DHS wealth index (and other indices).
Wu et al.[42]	2019	First, ResNet50 model pre-trained on ImageNet data, fine-tuned on night time light data. Second, linear r idge regression.	Input: first, LANDSAT 8 images. Second, the features from the ResNet50 model. Target: first, night time light data (Suomi NPP satellite). Second, Gross domestic product (GDP) and total retail sales of consumer goods (TRSCG).
Wu et al.[43]	2019	First, a ResNet50 (CNN) combined with a feature pyramid network (FPN). The ResNet-50 is pre-trained on ImageNet. Both networks are then trained (fine-tuned) to predict night-time light intensity categories and spectral index categories from the daytime satellite image.	Input: first, LANDSAT 8 images. Second, the features from the ResNet50 and the FPN. Target: first, night time light data (NPP-VIIRS) and spectral indices (NDVI, MNDWI, and NDBI). Second, per capita gross domestic product (PCGDP) from the Guizhou Bureau of Statistics.
Zhao et al.[44]	2019	RF. A CNN VGG-F, trained on ImageNet data, was fine-tuned to predict night time light classes (and thus learn features).	Input: night time lights from the VIIRS Cloud Mask–Outlier Removed (vcm–orm) annual composite NPP-VIIRS DNB data (NOAA/NCEI). Google Static Maps satellite images, zoom level 16 (high resolution). OSM. Land cover maps from the European Space Agency (ESA) Climate Change Initiative. Different features were computed from these. Target: the Wealth Index (WI) from the DHS.
Ayush et al.[45]	2020	Gradient descent boosting trees, but using deep neural networks (Yolov3) for object detection in the feature generation step.	Input: xView (DigitalGlobe) high-resolution satellite images. Features are extracted from these using Yolov3. Target: Living Standards Measurement Study for Uganda. Consumption expenditure.
Hofer et al.[46]	2020	ResNet-34 CNN models for images feature extraction and ridge regression for poverty target estimation.	Input: daytime satellite images from Landsat (15-m resolution) and Sentinel (10-m resolution). Night time satellite images from VIIRS. Target: poverty estimates from household surveys, and census data from the Philippines and Thailand.
Kondmann et al.[47]	2020	First, a ResNet 50 CNN, pre-trained on ImageNet and fine-tuned on night time lights. All layers are involved in the fine-tuning. Second, ridge regression on the features from the CNN.	Input: first, yearly composites of LANDSAT 7 satellite images. Second, the features from the CNN. Target: first, night time lights from DSMP-OLS. Second, the wealth index from the DHS data for Rwanda for the years 2005, 2010, and 2015. They test for the ability to predict the wealth index across time.

(*Continued on next page*)

Table 4. *Continued*

Reference	Year	Method	Data
Tan et al.[48]	2020	ResNet50 (CNN) for the first task. Ridge regression for the second task.	Input: first, Landsat 8 images. Second, the features from the ResNet50 network. Target: first, spectral index data (NDVI, MNDWI, and NDBI), and night time light data (Suomi NPP). Second, development indicators from the Chinese statistical yearbook data.
Yeh et al.[49]	2020	The ResNet-18 CNN model (v2, with preactivation). Pre-trained on ImageNet data. One network for daytime satellite images, one for night time.	Input: multispectral images from Landsat archives available on Google Earth Engine, and night-time lights images (VIIRS and DMSP). Target: the wealth index from the DHS data.
Ayush et al.[50]	2021	Gradient descent boosting trees. Deep neural networks (Yolov3) for object detection in the feature generation step. Reinforcement learning method for selecting low-resolution or high-resolution images.	Input: xView (DigitalGlobe) high-resolution satellite images and low-resolution Sentinel-2 satellite images. Target: LSMS for Uganda. Consumption expenditure.
Chi et al.[51]	2021	Gradient-boosted regression trees for wealth prediction and linear models for error prediction.	Input: features calculated from OSM, USGS, Facebook, VIIRS and satellite imagery from Digital Globe. Satellite image features are coming from pre-trained CNNs. Target: DHS "relative wealth index."
Engstrom et al.[52]	2021	Features are identified using a combination of CNN and classification of spectral and textural characteristics. In some cases (roads) is manual identification used. Finally, a linear LASSO model that uses the features as input.	Input: object and texture features derived from HSRI (High Spatial Resolution Imagery). Target: household estimates of per capita consumption imputed into the 2011 Census of population and housing (Sri Lanka).
Huang et al.[53]	2021	A deep learning model, Mask R-CNN, is tuned to detect and segment individual houses in satellite images. The color of the house roof is used to label its material (tin, thatched, or painted). The total roof area, and the fraction of high-quality roof areas out of the total roof area in the region, together with night time light intensity, are used as indicators of wealth. The Mask R-CNN model is first trained on two large public datasets, Common Objects in Context (COCO) and Open AI Tanzania, and then fine-tuned on a small set of samples from the study region in Kenya.	Input: first, high-resolution daytime satellite images from Google Static Maps, resolution 0.3 m. Second, the detected roofs, colors, and the night time lights. Target: first, manually segmented houses (for the small tuning set). Second, comparison with survey-based measures of economic well-being.
Jarry et al.[54]	2021	First, a VGG-16 CNN pre-trained on ImageNet and fine-tuned on night time lights. The VGG-16 network is followed by a fully connected layer. They also try to fine-tune with land-use as target, and to use a self-organized contrastive method to find good features. Second, ridge regression on the final features from the CNN.	Input: first, daytime satellite images from Google Static Map, with 2.5-m resolution. Second, the features from the CNN. Target: first, night time light data from the Earth Observation Group, extracted using Google Static Map. Second, poverty indicators from the World Bank's LSMS for Malawi.
Lee et al.[55]	2021	XGBoost (decision tree) for the feature based model. CNN for the satellite image-based model.	Input: OSM data, the VIIRS DNB night time lights dataset, day time satellite images (Google Static Map, zoom level 16), and the High-Resolution Settlement Layer (HRSL) datasets Target: the International Wealth Index (IWI) computed from the DHS data.

(*Continued on next page*)

Table 4. *Continued*

Reference	Year	Method	Data
Liu et al.[56]	2021	First, a VGG-16 CNN pre-trained on ImageNet and fine-tuned on night-time lights. The VGG-16 is modified with attention learning on the last layers. They also try non-fine-tuned VGG and a variational autoencoder (VAE) on the first step. Second, XGBoost is used as regression model on the features from the CNN.	Input: first, daytime satellite images from the PlanetScope Ortho Scene, with 5-m resolution. Second, features from the CNN encoded using principal component analysis (PCA); using the first 1–25 leading principal components. Target: first, night time lights from Earth Observation Group V1 annual composites, vcm version. Light intensities are grouped into three levels. Second, GDP for counties from the China Economic and Social Development Statistics Database, provided by China National Knowledge Infrastructure.
Ni et al.[57]	2021	CNNs. Four pre-defined architectures: VGG-Net, Inception-Net, ResNet and DenseNet. The two latter were also modified and tested as two new models.	Input: satellite images from Google Maps Static API. Fine-tuning with night time light imagery from DMSP-OLS satellite. Target: DHS data. Poverty index. Unclear what type.
Sako et al.[58]	2021	One approach with fine-tuning a CNN model on night time lights. The information is missing on the exact type of CNN. They also tried using the Yolov5 deep neural network object detector and adding the features from this to the night time lights predictor. The modeling from features to poverty data was tried with ridge regression, XGBoost, and RF.	Input: Sentinel 2 satellite images and high-resolution Google Earth images for a subset (2%) of the locations. The Yolov5 object detector was used to extract objects from the high-resolution images. Target: the first, fine-tuning step is done with VIIRS night time lights. In the second step, Philippines poverty data are used, which come from the Asian Development Bank.
Castro et al.[59]	2022	First, VGG16 CNN (transfer learning). Second, linear regression (combination of ridge and lasso).	Input: first, night time lights from VIIRS DNB), satellite images from Google Maps Static API, zoom level 16. Second, features computed from these images. Third, features from CNN models. Target: average income, GDP per capita, and water index in two Brazilian states.
Daoud et al.[60]	2022	Several CNN based models (e.g., ResNet-18, ResNet-34, ResNet-50, VGG-16).	Input: Landsat 7 daytime images (red, green, and blue soil reflectance bands, plus the panchromatic top-of-the-atmosphere band). Target: household data from the Indian censuses 2001 and 2011, and several targets from the Indian National Family Health Survey (NFHS).
Espín-Noboa et al.[61]	2022	A set of features are computed for each site: infrastructure from OSM and OpenCelliD, population and movement counts from Facebook data for good, audience reach estimates from Facebook Marketing, night light intensities from Google Earth Engine, and image featuresfrom a CNN trained to model the DHS IWI. (No information provided on if the CNN is pre-trained, or its structure.) The XGBoost algorithm is then used to model the IWI with different subsets of the features.	Input: 172 metadata-features (i.e., from OSM, OpenCelliD, and Facebook) plus 784 features from the CNN processing the daylight satellite image. Target: the IWI from DHS for Sierra Leone.
Tang et al.[62]	2022	First, a VGG-16 CNN, pre-trained on ImageNet and fine-tuned on night time light intensities. Second, a RF model built on the fine-tuned features.	Input: first, normalized difference vegetation index (NDVI) images, 16-day, from Google Earth Engine with a spatial resolution of 250 m. These are converted to monthly data. The NDVI approach is compared with the standard daytime satellite image approach, with images from Google Static Maps, zoom level 16. Target: first, night time lights from the Global DMSP-OLS Lights Time Series 1992–2013. Second, wealth index (and other indices) from DHS in Malawi, Nigeria, Rwanda, Tanzania, and Uganda.

Papers are listed chronologically and alphabetically (after first author).

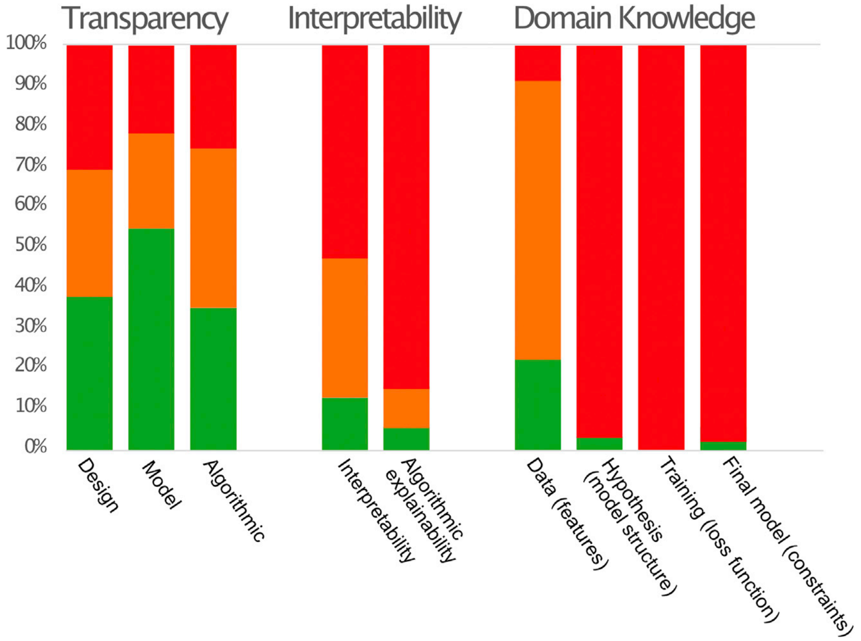

Figure 1. Classification of reviewed papers
Transparency bars: fractions of the reviewed papers that were classified as transparent (green), not transparent (red), and partly transparent (orange) with respect to the design, the models, and the algorithms used. Interpretability bars: fractions of the reviewed papers that were classified as interpretable (green), uninterpretable (red), and partly interpretable (orange), and corresponding fractions for algorithmic explainable. Domain knowledge bars: fractions of the reviewed papers that fulfilled the criteria for using domain knowledge (red), partly (orange), and no domain knowledge used (red), for data, hypothesis, training, and final model.

pixel values or be larger by masking out larger patches of the input image (occlusion).[69] It can highlight both negative and positive responses. A drawback is the computational requirement of perturbation methods, and it may be difficult to determine suitable perturbation schemes. Back-propagation-based methods for attribution maps are model specific and follow various forward and backward responses through the convolutional network's layers. Methods that compute gradients fall under this category, such as the Grad-CAM family.[70] Here, we also find DeepLIFT[71] and Deep SHAFT,[72] where the latter uses game theory-based SHAP values but is modified to work for deep-learning models.

Attribution maps are also relevant when satellite images predict human development indicators. The image regions that are most influential for the predictions can be highlighted. A challenge is, however, that it may be difficult to identify and quantify what regions are represented and how much they contribute. Domain knowledge will always be important when interpretable deep learning methods provide explanations.

METHODS

Our review method can be described as integrative rather than systematic,[73] and it originates from the sister publications of Xie et al.[16] and Jean et al.[11] These two publications illustrate the problems with reviewing this body of knowledge, mixing preprints, working papers, technical reports, peer-reviewed papers, and conference papers. The paper where Xie is the first author was published in the 30th AAAI Conference on Artificial Intelligence, in 2016, and appeared as an arXiv preprint in 2015. In 2016, the same group published a similar paper in *Science*. The latter paper builds the narrative for this review, as it represents a novelty in many aspects (covered in the introduction). Together, the two papers have been cited more than 1,500 times in Google Scholar (assessed July 27, 2022). The available literature in this research domain is presently not very large but growing rapidly, and we intend to cover it all. The literature body was collected by beginning with the papers we knew of in the domain and adding relevant papers in Google Scholar that cited them. Relevance was determined by reading the title and abstract, looking for the keywords "satellite," "machine learning," and "poverty" and "wealth." This yielded a rough set with about 100 papers that were read more thoroughly, and from which we selected 32 papers that fit our selection criteria:

1. The paper should cover poverty/wealth prediction, using survey data as a basis for the ground truth poverty/wealth estimates. It should be applicable to both urban and rural settings.
2. The method should use satellite images as the basis for at least some of the inputs (features).
3. The method should include deep neural networks.

Three papers that did fit these criteria were excluded because they were very low quality in terms of layout, spelling, and content, and were published in venues with a dubious reputation. In hindsight, all the 32 papers cite either Xie et al.[16] or Jean et al.[11] (except Xie et al. which was the first paper on the topic and, of course, did not cite itself).

The aim of the review is to provide an overview of how explainability is handled in this corpus of papers. Which aspects of XAI are covered well and which are not? Can these methods potentially be used to learn something new about poverty prediction? After all, it is perhaps surprising that the prediction accuracy on (asset) wealth is so high when all that is provided is a satellite image, with no information about the insides of buildings and a spatial resolution too low for some objects.

As far as possible, the structure provided by Roscher et al.[18] is used to organize the review. The presented models and analyses are evaluated with respect to the nine aspects described by Roscher et al.[18] as listed and described in Tables 1, 2, and 3: model transparency, design transparency, algorithmic transparency, interpretability, algorithmic explainability, using domain knowledge for feature design, using domain knowledge for generating the hypothesis, using domain knowledge in the training (loss function), and applying domain knowledge to the final model. How much the work in each paper fulfills these aspects is qualitatively grouped into three levels: well, somewhat, and poorly. In Figure 1, the three levels are represented with the three colors green, yellow, and red, respectively.

The categorization involves estimating boundary cases and can be illustrated with a few examples. One is the recent study by Lee and Braithwaite.[55] They use two ML models in their work: XGBoost and CNNs. Both result in well-defined deterministic functions. The parameters for training them are provided in the paper, and a researcher proficient in these methods should be able to reproduce them. Their work is labeled "green" for all three parts of transparency, although one could argue that it should perhaps be labeled "yellow" for the third part (algorithmic transparency). For the input to their models, they use well-defined sources, but on the output side (the label side) they both correct survey locations by hand and use a different WI than others. Therefore, their data should be made available for the work to be perfectly repeatable, but ethical aspects need to be considered. It is also unclear if their final prediction is a combination of the feature-based and the image-based models or if it is the output of only one of them. However, all their final models (parameters) should be possible to share without ethical issues and thus made available for others to explore for explainability, so their models are labeled transparent. Lee and Braithwaite[55] do not discuss how different features affect the output or what needs to be changed in an image for the output to change, so their work is labeled "red" for both interpretability and explainability. However, their work includes features believed to be important for the prediction, so we label it as "green" for domain-knowledge features. Domain knowledge is not used in other ways for the model building, e.g., for the cost function or to constrain the output, so the rest of the domain-knowledge aspects are labeled "red."

Another example is the study by Yeh et al.[49] where deep learning models are used to estimate asset wealth across approximately 20,000 African villages. The methods are well described from a transparency point of view, including details of obtaining the satellite data, good specification of the deep-learning models, and a good description of the model selection procedure such that other researchers can reproduce it. To further strengthen the transparency, both data and code are available for download (we did not test if the code worked but we checked that it is there). This paper, therefore, receives a "green" on all three transparency aspects. There are some anecdotal illustrations of the network responses to different images, but nothing more, and the label is, therefore, "yellow" and "red" for the interpretability and explainability.

Head et al.[30] present an interesting study where they model wealth and other indicators (for example, child weight and water accessibility) that could be related to wealth. They find that, whereas wealth can be well modeled, it is not as straightforward as the other indicators, which is surprising. This represents an exceptional case where domain knowledge is used to check if the models exhibit expected relationships, and this is the only paper that receives a "green" on the fourth domain—knowledge aspect.

The paper by Huang et al.[53] differs significantly from the others. It starts from the hypothesis that roof size and material reflects wealth, and from this builds an analysis of the effects of a financial support program. This is the only paper on the list that starts from a strong hypothesis and is also the only one to receive a "green" on the domain knowledge hypothesis.

RESULTS

Our review requires all papers to contain some aspect, including deep neural networks. The reviewed papers are listed in Table 4 with comments on the data and models used, and in chronological order. The work presented in each paper was evaluated with respect to the nine aspects listed above, and there were both clear and borderline cases, as illustrated in the examples above. The results should not be interpreted for each paper individually but as a result for the group of papers in this field.

The first part of the seven aspects deals with the transparency of the approaches. This relates to how well the work is documented, if the models can be repeated, and if the final models can be written down as functions in mathematical form. As Figure 1 shows, many papers do this well. However, far from all papers are written such that the work can be reproduced. The second part, interpretability and explainability, is a weak part in this field (and in many other fields too). Figure 1 shows that few researchers attempt to interpret their models, or even to illustrate what data leads to certain predictions. The explainability is even weaker; the explainable models tend to be simple decision trees or linear models. Very few researchers approach the issue of explaining the model prediction.

The third part of the seven aspects deals with domain knowledge: is domain knowledge used, e.g., to build the models, to select features, or to check if the final models satisfy expected constraints or behaviors. The use of domain knowledge for feature selection is common in the papers dealing with feature-based models. However, domain knowledge is not commonly used in other aspects of the modeling.

The purpose of this review is not to summarize all the experimental results in these papers, although some of them do reveal something about interpretability and explainability. One is the study mentioned above by Head et al.[30] which has been verified by Tingzon et al.[41]; the study shows that two indices related to wealth (e.g., wealth and access to water) cannot both be modeled well. Another is the work by Perez et al.[38] who received equally good predictions with low-resolution satellite images, indicating that the features that explain the results are perhaps not in the small details. A third result is in the work by Kondmann and Zhu,[47] where they conclude that the transfer-learning approach cannot be used to measure change in wealth over time, indicating that the features used by the transfer approach change very slowly.

There has, so far, been very little effort spent on interpreting and explaining why the transfer-learning approach works so (surprisingly) well. This is evident from the two interpretability bars in Figure 1.

DISCUSSION

Over the last decades, the dramatic success in ML has led to revitalization and progress in unexpected domains. Starting with the seminal papers of Jean et al.[11] and Xie et al.[16] it is now evident that some defined properties of poverty can be accurately estimated from combinations of satellite imagery and deep machine learning. Specifically, the fundamental questions of "where are the poor and how poor are they" could potentially be answered without launching a new wave of surveys. Recent research has accounted for some of the initial limitations that have been pointed out by Head et al.[30] and others, with generalizability being the most important. Lee and Braithwaite[55] have shown that it is possible to create a methodology that is generalizable to several countries. Chi et al.[51] have also presented a method that seems to generalize information concerning many countries. However, neither the approach by Lee and Braithwaite nor the approach by Chi et al. rests solely on the transfer-learning method and satellite images. While great progress has been achieved in a short time, several areas need attention, not least considering the reported lack of downstream applications of the methodology. We argue that explainability is essential to support research and applications, and explainability means more than just interpretability.

In this paper, we have investigated the readiness for this methodology to be used for scientific discovery, which is asking a lot, keeping in mind the recency of the research field. As a road map, we have used the requirements suggested by Roscher et al.[18] Our review of the field shows that the status of the three core elements of explainable ML (transparency, interpretability, and domain knowledge) is varied and does not completely fulfill the requirements set up for scientific insights and discoveries. Transparency matters are often well covered, meaning that most of the work is replicable and mathematically sound. Interpretability is not very well covered but, at the same time, interpretability accompanies the literature where recent efforts are directed and where we find state-of-the art research.[45,50] Furthermore, the use of domain knowledge, which is important to achieve scientific consistency, is not very well covered in the papers we have assessed. When scrutinizing research, as we do here, it is important to remember that the research field is in its infancy and has so far been advanced mainly by the technical community.

With its unprecedented spatial and temporal coverage, satellite data are increasingly used to measure various aspects of human welfare. The approach where outputs are benchmarked against survey data has produced some remarkable results but largely overlooked epistemological questions. Information evaluation is not well understood, specifically which imagery is physically capable of contributing to this specific domain. Head et al.[30] conclude that, if there is an insufficient signal in the image, “No matter how sophisticated our computational model, the model is destined to fail.” Therefore, some results that are intuitively inconsistent and difficult to explain. One such example observed by Head et al.[30] was the relative under-performance of models designed to predict access to drinking water. They expected the satellite-based features to capture proximity to bodies of water, which in turn might affect access to drinking water. However, an explainable AI approach here could perhaps shed light on this surprising finding.

Another discussion concerns the observed difference between stock-based and flow-based indices of poverty.[2,49] The relative under-performance of, for example, consumption expenditure is attributed to higher noise levels in training data than for assets. To resolve that issue, it is crucial to understand what it is in the image that triggers a certain response in the model. Indeed, the literature is unclear on the relationship between image features and surveyed features that impact output predictions, with few exceptions. It is plausible that the ML model is also picking up on some other features present in the image, such as the size and shape of arable land or road quality, and which are the properties that are known to be associated with welfare status. Complicating the matter further, we also know that there is evidence that much persistent poverty is place based and geographically determined,[49] meaning that there might be associations between landscape-specific properties and poverty to which the model is susceptible. For our concerns, sorting and ranking among impact features is probably the single most important future research direction. Herein lies one possibility for understanding something new about poverty and its determinants.

Understanding the interplay of different sensor characteristics (spatial and spectral resolutions), interactions with different physical environments, and the nature of ground truth is crucial. Satellite imagery is, in many aspects, about what we see is what we get. A textbook rule-of-thumb is that for an object to be observable, it should be covered with a minimum of four pixels. Image enhancement techniques and combining spectral information in creative ways can bring out extra detail, and high-contrast objects are also more likely to be observable. What could we expect to observe if we consider the commonly used 2.5-m satellite image (scale level 16 in Google API), imagining it centered over one of the DHS rural villages (provided displacement of coordinates)? The dominating objects would be buildings, roads, agricultural fields, and forest patches, but also the overall spatial organization of the village. In other words, many features are not accounted for in WI design. It would be difficult to identify cars (but maybe possible), bicycles, electricity poles, cell phones, television sets, farm animals, although they are all features of the WI. Results seem to depend very little on resolution and even on time, so why is it unlikely that the reason for the outcome depends on the image resolution?

Is this consistent with domain knowledge? This information would shed light on the important relationship between geographically determined poverty and other forms of poverty. The recent studies by Ayush et al.[45,50] represent interesting starting points. Here, objects were detected in the images using object detection methods as a first step. A feature vector was then constructed using the counts of different objects, augmented with confidence and size estimates. Using a tree-based model, these feature vectors were then used to model a poverty index. This approach provides an increased level of explainability due to features that are directly interpretable and models that inherently provide some explanations of how features affect the outcome. There is still room for improvement in terms of expanding the set of objects to detect or adding more abstract features, such as landscape characteristics, to increase model explainability further. The advantage of a traditional equation, showing both which terms there are in the equation and what factors there are in front of each term, is obvious.

An example to illustrate this is the Cobb Douglas production function commonly found in economics of small firms but also in agricultural production (from Neumann et al.[74]),

$$\ln(q) = \sum_i \beta_i \ln(X_i) + v - u,$$

where $\ln(q)$ is the logarithm of the production of the grid cell in question, X_i are the different production inputs for that cell, β_i are unknown parameters to be estimated, and v is a random error to account for statistical noise. Further deviations are due to inefficiencies (u) for that grid cell. The function for crop production (cp) could be

$$\begin{aligned}\ln(cp) = {} & \beta_0 + \beta_1 \ln(\mathit{temperature}) + \beta_2 \ln(\mathit{precipitation}) \\ & + \beta_3 \ln(\mathit{par}) + \beta_4 \ln(\mathit{soilfertility}) + v - u,\end{aligned}$$

with the most important growth defining factors according to theory inserted; temperature, precipitation, photo-synthetically active radiation (par), and a soil fertility constant (all values estimated for the particular grid cell studied). For the inefficiency (u), influences of land management, labor force, general accessibility, and market access are considered important:

$$\begin{aligned}u = {} & \delta_1(\mathit{irrigation}) + \delta_2(\mathit{slope}) + \delta_3(\mathit{agripopulation}) \\ & + \delta_4(\mathit{access}) + \delta_5(\mathit{market}).\end{aligned}$$

A model of this kind provides grounds for evaluating the effect of changes in explanatory variables on $\ln(q)$ and deviations from expected levels (the inefficiency function). It would be relevant to extract a similar explanation from the ML- and satellite image-based models for poverty estimation. What features in the satellite image form the basis for the poverty estimate? Are there expected base level and place-specific inefficiencies relating to the particular grid cell?

For a period, the scholars have been occupied with increasing model performance to very good levels compared with survey data. The next natural step is working toward increased levels of explainability to explain how features affect the outcome. From our perspective, which marks the start of learning something new about poverty and its distributional properties and would be a venue for different domain experts to work together. One approach that could be useful and a complement to the contribution of Ayush et al.[45,50] is to work with heat maps. A heatmap represents coefficients to visualize the strength of correlation among variables and could add to the discussion on what is in the image that contributes to a certain poverty score. However, having an explainable model requires not only knowing what features or objects that are important for an estimate but also describing how these features affect the estimate. Eventually, it will be possible to ask the important questions, such as: "What does positive change in poor societies look like and how is it achieved?"[75] However, is it likely to be answered from imagery alone?

ACKNOWLEDGMENTS

The authors would like to thank financial support during the project from the Swedish Research Council 2019-04253 and Riksbankens Jubileumsfond MXM19-1104:1. We would also like to thank the reviewers for their insightful comments on the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. McBride, L., Barrett, C.B., Browne, C., Hu, L., Liu, Y., Matteson, D.S., Sun, Y., and Wen, J. (2021). Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. Appl. Econ. Perspect. Pol. *44*. https://doi.org/10.1002/aepp.13175.

2. Burke, M., Driscoll, A., Lobell, D.B., and Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. Science *371*, 1–12. https://doi.org/10.1126/science.abe8628.

3. Espey, J., Swanson, E., Badiee, S., Christensen, Z., Fischer, A., and Levy, M. (2015). Data for development: a needs assessment for SDG monitoring and statistical capacity development (the Sustainable Development Solutions Network (SDSN), United Nations). Technical Report. https://sustainabledevelopment.un.org/content/documents/2017Data-for-Development-Full-Report.pdf.

4. Jerven, M. (2017). How much will a data revolution in development cost? Forum Dev. Stud. *44*, 31–50. https://doi.org/10.1080/08039410.2016.1260050.

5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

6. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.

7. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer International Publishing), pp. 630–645. https://doi.org/10.1007/978-3-319-46493-0_38.

8. Ho-Phuoc, T. (2018). Cifar10 to compare visual recognition performance between deep neural networks and humans. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.07270.

9. Mikami, H., Suganuma, H., U-chupala, P., Tanaka, Y., and Kageyama, Y. (2018). Massively distributed SGD: ImageNet/ResNet-50 training in a flash. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.05233.

10. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. In Artificial Neural Networks and Machine Learning – ICANN 2018, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, eds. (Springer International Publishing), pp. 270–279. https://doi.org/10.1007/978-3-030-01424-7_27.

11. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science *353*, 790–794. https://doi.org/10.1126/science.aaf7894.

12. Atzberger, C. (2013). Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs. Rem. Sens. *5*, 949–981. https://doi.org/10.3390/rs5020949.

13. Weiss, M., Jacob, F., and Duveiller, G. (2020). Remote sensing for agricultural applications: a meta-review. Remote Sens. Environ. *236*, 111402. https://doi.org/10.1016/j.rse.2019.111402.

14. Archila Bustos, M.F., Hall, O., Niedomysl, T., and Ernstson, U. (2020). A pixel level evaluation of five multitemporal global gridded population datasets: a case study in Sweden, 1990–2015. Popul. Environ. *42*, 255–277. https://doi.org/10.1007/s11111-020-00360-8.

15. Kuffer, M., Pfeffer, K., and Sliuzas, R. (2016). Slums from space—15 years of slum mapping using remote sensing. Rem. Sens. *8*, 455. https://doi.org/10.3390/rs8060455.

16. Xie, M., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2016). Transfer learning from deep features for remote sensing and poverty mapping. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI'16 (AAAI Press). 3929—3935. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12196/12181.

17. Blumenstock, J. (2020). Machine learning can help get covid-19 aid to those who need it most. Nature. https://doi.org/10.1038/d41586-020-01393-7.

18. Roscher, R., Bohn, B., Duarte, M.F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. IEEE Access *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

19. Council, N.R. (1998). People and Pixels: Linking Remote Sensing and Social Science (The National Academies Press). https://doi.org/10.17226/5963.

20. Hall, O. (2010). Remote sensing in social science research. Open Rem. Sens. J. *3*, 1–16. https://doi.org/10.2174/1875413901003010001.

21. Longley, P.A. (2002). Geographical information systems: will developments in urban remote sensing and gis lead to 'better'urban geography? Prog. Hum. Geogr. *26*, 231–239. https://doi.org/10.1191/0309132502ph366pr.

22. Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W., Davis, E.R., and Davis, C.W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. Int. J. Rem. Sens. *18*, 1373–1379. https://doi.org/10.1080/014311697218485.

23. Henderson, J.V., Storeygard, A., and Weil, D.N. (2012). Measuring economic growth from outer space. Am. Econ. Rev. *102*, 994–1028. https://doi.org/10.1257/aer.102.2.994.

24. Chen, X., and Nordhaus, W.D. (2011). Using luminosity data as a proxy for economic statistics. Proc. Natl. Acad. Sci. USA *108*, 8589–8594. https://doi.org/10.1073/pnas.1017031108.

25. Keola, S., Andersson, M., and Hall, O. (2015). Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. World Dev. *66*, 322–334. https://doi.org/10.1016/j.worlddev.2014.08.017.

26. Noor, A.M., Alegana, V.A., Gething, P.W., Tatem, A.J., and Snow, R.W. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. Popul. Health Metr. *6*, 5–13. https://doi.org/10.1186/1478-7954-6-5.

27. Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. (2015). Night-time light data: a good proxy measure for economic activity? PLoS One *10*, e0139779. https://doi.org/10.1371/journal.pone.0139779.

28. Andersson, M., Hall, O., and Archila, M.F. (2019). How data-poor countries remain data poor: underestimation of human settlements in Burkina Faso as observed from nighttime light data. ISPRS Int. J. Geo-Inf. *8*, 498. https://doi.org/10.3390/ijgi8110498.

29. Chen, X., and Nordhaus, W. (2015). A test of the new viirs lights data set: population and economic output in africa. Rem. Sens. *7*, 4937–4947. https://doi.org/10.3390/rs70404937.

30. Head, A., Manguin, M., Tran, N., and Blumenstock, J.E. (2017). Can human development be measured with satellite imagery? In Proceedings of the Ninth International Conference on Information and Communication Technologies and Development. ICTD '17 (Association for Computing Machinery), pp. 1–11. https://doi.org/10.1145/3136560.3136576.

31. Ravallion, M. (2020). On measuring global poverty. Annu. Rev. Econom. *12*, 167–188. https://doi.org/10.1146/annurev-economics-081919-022924.

32. Zhou, Y., and Liu, Y. (2022). The geography of poverty: review and research prospects. J. Rural Stud. *93*, 408–416. https://doi.org/10.1016/j.jrurstud.2019.01.008.

33. Rutstein, S.O., and Staveteig, S. (2014). Making the Demographic and Health Surveys Wealth Index Comparable, *9* (ICF International Rockville).

34. Chen, D. (2017). Temporal Poverty Prediction Using Satellite Imagery (Stanford University, Department of Computer Science). Technical Report. http://cs231n.stanford.edu/reports/2017/pdfs/552.pdf.

35. Kim, J.H., Xie, M., Jean, N., and Ermon, S. (2016). Incorporating Spatial Context and Fine-Grained Detail from Satellite Imagery to Predict Poverty (Stanford University, Department of Computer Science). Technical Report. https://doi.org/10.13140/RG.2.2.27604.60803.

36. Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., and Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.06323.

37. Irvin, J., Laird, D., and Rajpurkar, P. (2017). Using Satellite Imagery to Predict Health (Stanford University, Department of Computer Science). Technical Report. http://cs231n.stanford.edu/reports/2017/pdfs/559.pdf.

38. Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D., and Ermon, S. (2017). Poverty prediction with public landsat 7 satellite imagery and machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.03654.

39. Pandey, S., Agarwal, T., and Krishnan, N.C. (2018). Multi-task deep learning for predicting poverty from satellite images. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI Press), pp. 7793–7798. https://doi.org/10.1609/aaai.v32i1.11416.

40. Perez, A., Ganguli, S., Ermon, S., Azzari, G., Burke, M., and Lobell, D. (2019). Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. Preprint at arXiv. https://doi.org/10.48550/arXiv.1902.11110.

41. Tingzon, I., Orden, A., Go, K.T., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M., and Kim, D. (2019). Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. *4219*, 425–431. https://doi.org/10.5194/isprs-archives-XLII-4-W19-425-2019.

42. Wu, P., and Tan, Y. (2019a). Estimation of economic indicators using residual neural network ResNet50. In 2019 International Conference on Data Mining Workshops (ICDMW), pp. 206–209. https://doi.org/10.1109/ICDMW.2019.00039.

43. Wu, P., and Tan, Y. (2019b). Estimation of poverty based on remote sensing image and convolutional neural network. Adv. Rem. Sens. *08*, 89–98. https://doi.org/10.4236/ars.2019.84006.

44. Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C., and Wu, J. (2019). Estimation of poverty using random forest regression with multi-source data: a case study in Bangladesh. Rem. Sens. *11*, 375. https://doi.org/10.3390/rs11040375.

45. Ayush, K., Uzkent, B., Burke, M., Lobell, D., and Ermon, S. (2020). Generating interpretable poverty maps using object detection in satellite images. Preprint at arXiv. https://doi.org/10.48550/arXiv.2002.01612.

46. Hofer, M., Sako, T., Martinez, A., Jr., Addawe, M., Bulan, J., Durante, R.L., and Martillan, M. (2020). Applying Artificial Intelligence on Satellite Imagery to Compile Granular Poverty Statistics (Asian Development Bank). Technical Report ADB Economics Working Paper Series No. 629. https://doi.org/10.22617/WPS200432-2.

47. Kondmann, L., and Zhu, X.X. (2020). Measuring changes in poverty with deep learning and satellite images. In International Conference on Learning Representations (ICLR) 2020, Practical ML for Developing Countries Workshop, pp. 1–6. https://elib.dlr.de/137108/2/camera_ready.pdf.

48. Tan, Y., Wu, P., Zhou, G., Li, Y., and Bai, B. (2020). Combining residual neural networks and feature pyramid networks to estimate poverty using multisource remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. *13*, 553–565. https://doi.org/10.1109/JSTARS.2020.2968468.

49. Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nat. Commun. *11*, 2583. https://doi.org/10.1038/s41467-020-16185-w.

50. Ayush, K., Uzkent, B., Tanmay, K., Burke, M., Lobell, D., and Ermon, S. (2021). Efficient poverty mapping from high resolution remote sensing images. In Proceedings of the AAAI Conference on Artificial Intelligence, *35*, pp. 12–20. https://ojs.aaai.org/index.php/AAAI/article/view/16072.

51. Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J.E. (2022). Microestimates of wealth for all low- and middle-income countries. Proc. Natl. Acad. Sci. USA *119*. e2113658119. https://doi.org/10.1073/pnas.2113658119.

52. Engstrom, R., Hersh, J., and Newhouse, D. (2021). Poverty from space: using high resolution satellite imagery for estimating economic well-being. World Bank Econ. Rev. *36*, 382–412. https://doi.org/10.1093/wber/lhab015.

53. Huang, L.Y., Hsiang, S.M., and Gonzalez-Navarro, M. (2021). Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-poverty Programs (National Bureau of Economic Research). Technical Report Working Paper 29105. https://doi.org/10.3386/w29105.

54. Jarry, R., Chaumont, M., Berti-Équille, L., and Subsol, G. (2021). Assessment of CNN-based methods for poverty estimation from satellite images. In PRRS 2021 – 11th IAPR International Workshop on Pattern Recognition in Remote Sensing, pp. 550–565. https://doi.org/10.1007/978-3-030-68787-8_40.

55. Lee, K., and Braithwaite, J. (2020). High-resolution poverty maps in sub-Saharan africa. Preprint at arXiv. https://doi.org/10.48550/arXiv.2009.00544.

56. Liu, H., He, X., Bai, Y., Liu, X., Wu, Y., Zhao, Y., and Yang, H. (2021). Nightlight as a proxy of economic indicators: fine-grained GDP inference around Chinese mainland via attention-augmented CNN from daytime satellite imagery. Rem. Sens. *13*, 2067–5736. https://doi.org/10.3390/rs13112067.

57. Ni, Y., Li, X., Ye, Y., Li, Y., Li, C., and Chu, D. (2021). An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction. IEEE Geosci. Remote Sens. Lett. *18*, 1545–1549. https://doi.org/10.1109/LGRS.2020.3006019.

58. Sako, T., and Martinez, A.J.M. (2021). Seeing poverty from space, how much can it be tuned?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.14700.

59. Castro, D.A., and Álvarez, M.A. (2022). Predicting socioeconomic indicators using transfer learning on imagery data: an application in Brazil. Geojournal, 1–22. https://doi.org/10.1007/s10708-022-10618-3.

60. Daoud, A., Jordan, F., Sharma, M., Johansson, F., Dubhashi, D., Paul, S., and Banerjee, S. (2022). Using satellites and artificial intelligence to measure health and material-living standards in India. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.00109.

61. Espín-Noboa, L., Kertész, J., and Karsai, M. (2022). Challenges of inferring high-resolution poverty maps with multimodal data. In International Conference on Learning Representations (ICLR) 2022, Practical ML for Developing Countries Workshop, pp. 1–8. https://pml4dc.github.io/iclr2022/pdf/PML4DC_ICLR2022_17.pdf.

62. Tang, B., Liu, Y., and Matteson, D.S. (2022). Predicting poverty with vegetation index. Appl. Econ. Perspect. Pol. *44*, 930–945. https://doi.org/10.1002/aepp.13221.

63. Lipton, Z.C. (2018). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue *16*, 31–57. https://doi.org/10.1145/3236386.3241340.

64. Montavon, G., Samek, W., and Müller, K.R. (2018). Methods for interpreting and understanding deep neural networks. Digit. Signal Process. *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.

65. Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

66. von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., et al. (2021).

Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans. Knowl. Data Eng. https://doi.org/10.1109/TKDE.2021.3079836.

67. van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., and Viergever, M.A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Med. Image Anal. *79*, 102470. https://doi.org/10.1016/j.media.2022.102470.

68. Gulum, M.A., Trombley, C.M., and Kantardzic, M. (2021). A review of explainable deep learning cancer detection models in medical imaging. Appl. Sci. *11*, 4573. https://doi.org/10.3390/app11104573.

69. Zeiler, M.D., and Fergus, R. (2014). Visualizing and understanding convolutional networks. In Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds. (Springer International Publishing), pp. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

70. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. https://doi.org/10.1109/ICCV.2017.74.

71. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17 (JMLR.org), pp. 3145–3153. https://doi.org/10.5555/3305890.3306006.

72. Chen, H., Lundberg, S., and Lee, S.I. (2019). Explaining models by propagating shapley values of local components. Preprint at arXiv. https://doi.org/10.48550/arXiv.1911.11888.

73. Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines. J. Bus. Res. *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039.

74. Neumann, K., Verburg, P.H., Stehfest, E., and Müller, C. (2010). The yield gap of global grain production: a spatial analysis. Agric. Syst. *103*, 316–326. https://doi.org/10.1016/j.agsy.2010.02.004.

75. Östberg, W., Howland, O., Mduma, J., and Brockington, D. (2018). Tracing improving livelihoods in rural Africa using local measures of wealth: a case study from central Tanzania, 1991–2016. Land *7*, 44. https://doi.org/10.3390/land7020044.