Predicting Maternal and Infant Health Outcomes in Western Africa using Satellite Images

Sauren Khosla
Department of Computer Science
Stanford University
sauren@stanford.edu

Benjamin Wittenbrink
Department of Computer Science
Stanford University
witten@stanford.edu

Caroline Zanze
Department of Computer Science
Stanford University
ckzanze@stanford.edu

Abstract

Accurately predicting which areas require additional attention and assistance in the medical world is an important and critical task for augmenting global health outcomes. Doing so with limited data is of particular interest, as the least supported areas are least likely to have infrastructure in place to facilitate the acquisition of additional healthcare resources. In this paper, we use satellite images from NASA's Landsat database to predict malnourishment (operationalized by mean BMI) and child mortality rates. We assess the performance of a variety of computer vision models, task-agnostic MOSAIKS feature vectors, and metadata-enhanced fusion models on ordinal discretizations of our outcome variables. Our best model, a fine-tuned Vision Transformer pre-trained on ImageNet, achieves a balanced accuracy of 66.7%, doubling random chance, on a CDC-derived ordinal encoding of Body Mass Index (BMI), and an accuracy of 50.3%, a 51% improvement over the random baseline, on an ordinal encoding of child mortality rates. We provide suggestive evidence that the model attends to sociologically relevant aspects of the images, such as road networks and city/village layouts. Future work should entail further enhancements of the most promising Vision Transformer model, further qualitative analysis and visualization of the model's predictions, and extensions to other measures capturing a region's health outcomes.

1 Introduction

The main contribution of this paper is to utilize satellite images and satellite-derived variables to predict basic indicators of maternal and child health (MCH) in western Africa at the village-level. Such indicators are primarily obtained from nationally representative household surveys such as the Demographic Health Survey (DHS). These surveys have many limitations; they are expensive in terms of pecuniary cost and human capital and consequently can suffer from limited temporal resolution and representativeness, as only a small minority of individuals can be selected for interviews.

The accessibility and coverage of satellite imagery present a promising opportunity to take advantage of recent advances in computer vision and machine learning to build a model to provide real-time estimates of MCH at a village level. Intuitively, satellite images contain information about a variety of causes of MCH outcomes. These causes include wealth and poverty of a village; accessibility of fresh water and nutrients; interconnectedness of a village within the broader region; microclimate and agriculture; and infrastructure quality. Existing research demonstrates that satellite imagery can be used to reliably predict a number of these features.

We focus on predicting two basic MCH indicators at the village-level: mean Body Mass Index (BMI) and under-five mortality (UFM) rate. Both of these variables are continuous and thus most naturally suited to a regression problem. However, to make the problem more tractable, we discretize the outcomes to form a classification task.

2 Related Work

Although still a nascent field, deep learning models have been increasingly used to predict economic data from satellite images. Engstrom et al. used object detection to extract features (e.g. number of buildings, type of agriculture, and roof material) from satellite images of Sri Lanka, which they subsequently fed into a simple linear regression model to predict poverty in neighboring areas [1]. Jean et al. used transfer learning to produce a CNN which predicts consumption expenditure and asset wealth from satellite images of African countries. The authors used a CNN pretrained on ImageNet to extract low-level image features, which they fine-tuned to predict nighttime light intensities corresponding to input daytime satellite images. The learned nighttime light intensity served as a proxy for GDP and was used to train the final model [2]. Yeh et al. also used nightlights to estimate wealth in Nigeria, but they used nightlight images as inputs to the model instead of employing transfer learning to predict them indirectly [3]. Finally, Huang et al. used satellite imagery to evaluate the impact of anti-poverty programs using the Mask R-CNN model for instance segmentation of buildings. For each building instance, the size of the building footprint and the roof material were extracted and used as proxies for housing quality, which was ultimately used as a proxy for household wealth. The authors concluded that deep learning can complement and in some cases substitute in-person survey data [4]. The successes of these models in predicting economic data from satellite imagery is promising for our task, for economic conditions are certainly determinants of a population's health.

Infrastructure quality and agricultural production may also be useful health indicators. Ohsri et al. used an 18-layer ResNet to predict infrastructure quality of electricity, sewerage, piped water, and roads in developing African regions from satellite images [5]. As for agricultural production, You et al. study sought to predict soybean crop yields in the United States. They trained both a CNN and a LSTM and added a Gaussian Process layer atop these networks to account for spatio-temporal dependencies between data points [6].

There also exist precedents for using satellite imagery to directly predict health data. Bibault et al. trained a model to predict cancer prevalence using satellite imagery and cancer prevalence estimates from the CDC's 500 (U.S.) Cities Project. They extracted image features from ResNet and then fed them into an elastic net [7]. Levy et al. sought to predict county-level mortality rates in the US from satellite images by fine-tuning ResNet. They used Shapley Additive Feature Explanations (SHAP) to identify relevant features in satellite images across their test counties and found that lower mortality rate is associated with environmental features such as sidewalks and hiking trails [8]. Additionally, Rolf et al. obtain a task-agnostic high-dimensional feature representation of satellite images with their Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS) model. The authors show that these features generalize across 9 diverse prediction tasks such as forest cover, house price, and road length. Furthermore, while achieving similar accuracy to deep learning networks, it expends a fraction of the computational cost [9].

The study most similar to our task was conducted by Adyasha Maharana and Elaine Nsoesie at The University of Washington. They predicted the prevalence of adult obesity from satellite images using the VGG-CNN-F network (to extract gradients, edges, and patterns to aid in object detection) fed into an ElasticNet regression model. They found that physical characteristics of a neighborhood, such as the presence of parks, highways, green streets, and crosswalks, are associated with variations in obesity prevalence [10]. Notably, their satellite images were limited to 7 large U.S. cities, whereas we train our model on more developing areas.

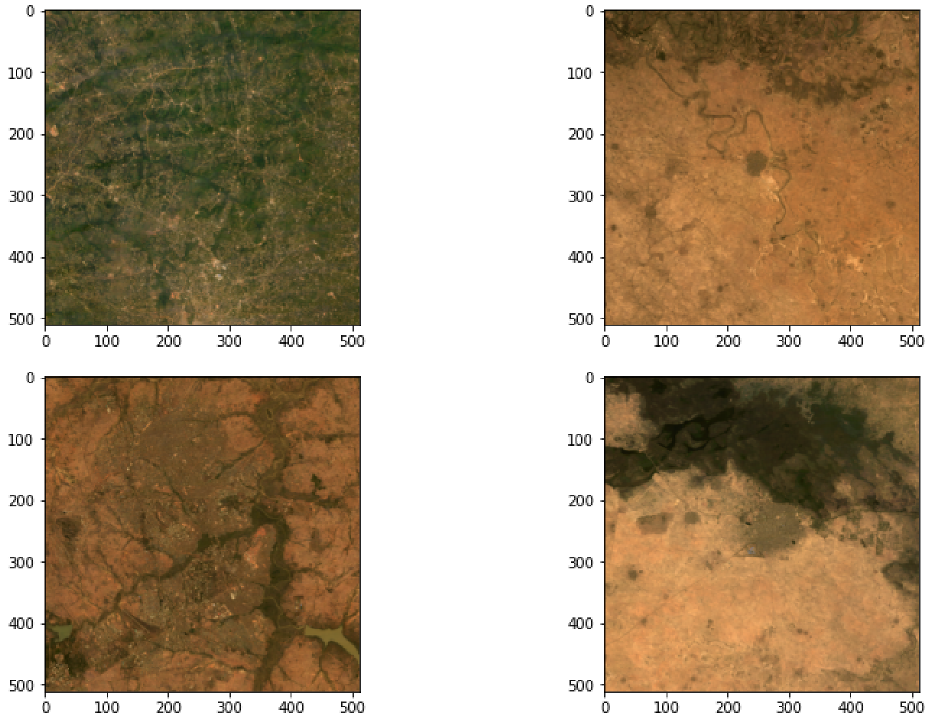
3 Data

Our data set consists of NASA's Landsat satellite images combined with Demographic and Health Survey (DHS) data as well as purely tabular DHS data. The satellite data entail 157 folders, each representing a country and the year the images were taken. Each folder has an average of 750 511x511 images (though more recent years have more images). Specifically, the data are stored in .tfrecords, which contain a byte-string for the data-payload, the data-length, and CRC-32C hashes for integrity checking. Each data payload contains one 511x511 satellite image which represents 6.75 square kilometers. Each image has channels for red, green, blue, NIR, SWIR, country code, image number, DHS-ID, longitude and latitude of the cluster, and the year the image was taken. The DHS-ID contains the year the survey was taken as well as the DHS cluster code, which represents the groupings of households that participated in the survey. NIR and SWIR encode infrared information, but we do not use these channels as inputs to our model. We initially normalize the raw RGB values of each image according to a minimum value of zero and a maximum of 255 (without this normalization the images appear very dark).

Our labels come from DHS surveys, which can be connected to our satellite images using the DHS-ID. Specifically, for each DHS-ID, we have the DHS year, DHS cluster, latitude and longitude of the cluster, mean BMI, unmet need rate, under five mortality (UFM) rate, skilled birth attendant rate, stunted rate, fully vaccinated children rate, and an indicator for urban or rural.

For the purposes of this study, we restrict our focus to predicting the mean BMI and UFM rate for each image. Moreover, for computational reasons, we restrict our main sample to a set of 14 western African countries, taking the most recent satellite imagery for each country.[1] All but one of the country images were taken since 2010. Figure 1 depicts two images associated with relatively higher mean BMIs (greater than 25) and two with relatively lower mean BMIs (less than 20), all from Nigeria and taken in 2018. Figure 2 plots the geographic distribution of our sample, where each point indicates an image and is colored by our outcome variable. The high-level spatial distribution of these values conform to our *a priori* understanding: areas along the coast are more affluent and thus have higher BMIs and lower mortality rates whereas areas near to and in the Sahara are substantially poorer and thus have lower BMIs and higher mortality rates. Also note that child mortality rate varies greatly across both urban and rural environments (see Appendix Figure A1).

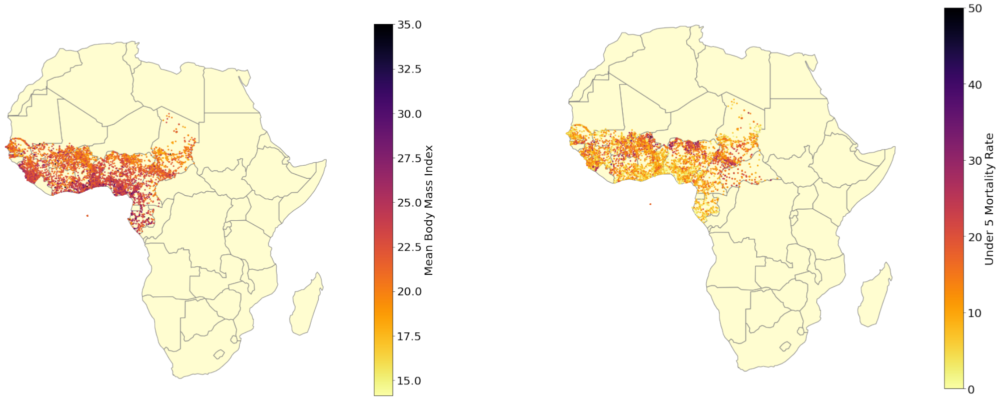
[1] In particular, we select the following country-year pairs: Burkina Faso (2010), Benin (2012), Central African Republic (1994), Ivory Coast (2012), Cameroon (2011), Gabon (2012), Ghana (2014), Guinea (2018), Mali (2012), Nigeria (2018), Sierra Leone (2019), Senegal (2010), Chad (2014), and Togo (2013).



(a) High BMI Images (> 25)

(b) Low BMI Images (< 20)

Figure 1: Examples of satellite image data, Nigeria 2018



(a) Colored by body mass index

(b) Colored by under five mortality rate

Figure 2: Geographic distribution of images, colored by outcome

4 Approach

4.1 Outcome Construction

Mean BMI and UMF rate are continuous variables and are thus most naturally suited for a regression task. However, we simplified the problem into a classification task, which proved to be sufficiently challenging and is medically motivated, as it is most relevant to public health to detect and understand values at the extremes of the distribution.

Figure 3 depicts the continuous distribution of Mean BMI scores (left) and UFM rates (right) in our sample. The Mean BMI distribution resembles a Gaussian centered around 23 with a slightly longer

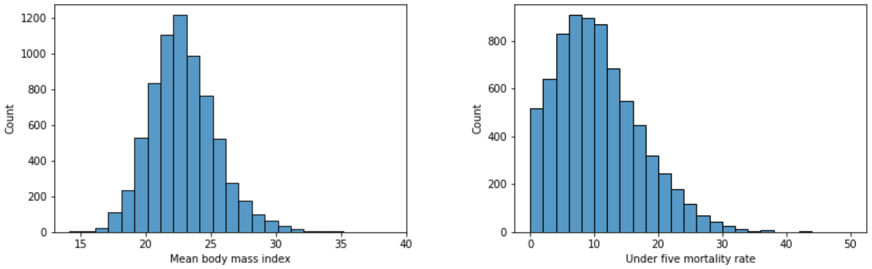


Figure 3: Histogram of BMI scores, UFM rates for western Africa satellite images

right tail. One natural discretization of this distribution might be to bin the sequence into quintiles. Thus, first we group the images into the following quintiles, which produces the following cuts: 20.8 and below, 20.8 to 22.1, 22.1 to 23.3, 23.3 to 24.8, and 24.8 and above. An alternative approach would be to discretize the data according to commonly define medical ranges. For example, the CDC classifies any BMI under 18.5 as underweight, BMIs between 18.5 and 24.9 as healthy, between 25 and 29.9 as overweight, and 30 or above as obese. As only 15 images fall into the obese and underweight categories each, we coarsen the mapping into the following groups: 19.9 and below, between 20 and 24.9, and 25 and above.

We follow a similar procedure for UFM rates. First, we discretize according to quintiles, which leads to the bins: 4.8 and below, 4.8 to 8.1, 8.1 to 11.4, 11.4 to 16, and 16 and above. In addition, we specify a more intuitive ‘severity" binning, corresponding to 10 and below (“low"), 10 to 20 (“medium"), and 20 and above (“high").

4.2 Models

4.2.1 Baselines

As our baselines, we compare our preferred models with naive, simple alternatives. First, we specify a logistic regression on image data. Second, we use a basic two layer convolutional network where each layer corresponds to a block of batch normalization (BN), 3×3 convolution, ReLU, and 2×2 max pooling. This is then inputted to a dropout layer and a final fully connected layer.

We also specify a baseline using the MOSAIKS model to generate embeddings for our images and then fed them into a classification algorithm. In particular, the MOSAIKS model begins with a set of images $\{\mathbf{I}_\ell\}_{\ell=1}^N$, each of which is centered at locations indexed by $\ell = \{1, \ldots, N\}$. The MOSAIKS model then generates task-agnostic feature vectors $\mathbf{x}(\mathbf{I}_\ell)$ for each satellite image $\mathbf{I}_\ell$ by convolving an $M \times M \times S$ “patch", P_k, across the entire image, where M is the width and height of the patch in units of pixels and S is number of spectral bands. In each convolution step, the inner product of the patch and a $M \times M \times S$ sub-image region is taken, and a ReLU activation function with bias $b_k = 1$ is applied. Each patch is a randomly sampled sub-image from the training images $\{\mathbf{I}_\ell\}_{\ell=1}^N$ [9].

4.2.2 Vision Transformer (ViT)

Our main model is a Vision Transformer (ViT), first introduced by [11] as an adaption of the seminal Transformer architecture by [12] from natural language processing. We provide a visual overview of this model in Figure 4 using an illustration from [11].

The standard Transformer is designed for text data and thus accepts a one-dimensional input of word token embeddings. To adapt a Transformer to image data, the ViT is modified to take in a transformed two-dimensional input. In particular, for a three-dimensional image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ where H and W represent the height and width respectively and C the number of channels, ViT flattens it into a sequence of image patches $\boldsymbol{x}_p \in \mathbb{R}^{(HW/P^2) \times (P^2 C)}$ such that P represents the desired resolution of each image patch. Consequently, $N = HW/P^2$. These patches are then flattened using a linear

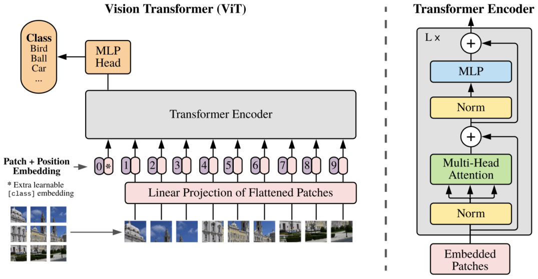


Figure 4: Vision Transformer (ViT) Architecture [11]

projection into $N \times D$ space to obtain a feature extraction $\boldsymbol{z}_0$:

$$\boldsymbol{z}_0 = [\boldsymbol{x}_{\text{class}};\ \boldsymbol{x}_p^1\boldsymbol{E};\ \dots;\ \boldsymbol{x}_p^N\boldsymbol{E}] + \boldsymbol{E}_{pos} \tag{1}$$

for $\boldsymbol{E} \in \mathbb{R}^{(P^2C)\times D}$ and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$, a matrix of positional encodings. We denote $\boldsymbol{z}_0^0 = \boldsymbol{x}_{\text{class}}$.

This value is then taken as input to the model encoder, consisting of alternating layers of multi-headed self-attention (MSA) and multi-layer perceptron (MLP) blocks, which contain two layers and a Gaussian Error Linear Unit (GELU) activation function each. The MSA layer is defined as in [12] using scaled dot product attention, that is for queries $\boldsymbol{q}$, keys $\boldsymbol{k}$, and values $\boldsymbol{v}$:

$$\begin{aligned} [\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] &= \boldsymbol{z}\boldsymbol{U}_{qkv} & \boldsymbol{U}_{qkv} &\in \mathbb{R}^{D\times 3D_k} \\ \boldsymbol{A} &= \text{softmax}(\frac{\boldsymbol{q}\boldsymbol{k}^T}{\sqrt{D_k}}) & \boldsymbol{A} &\in \mathbb{R}^{N\times N} \\ SA(\boldsymbol{z}) &= \boldsymbol{A}\boldsymbol{v} \end{aligned} \tag{2}$$

where $D_k = D/k$. Intuitively, $\boldsymbol{A}$ captures the pair-wise similarity between elements. Then MSA with k attention heads is just a concatenation of these k outputs linearly projected back into D by a matrix $\boldsymbol{U}_{msa}^{kD_k\times D}$. Before each block, a layernorm (LN) is applied and residual connections are introduced after each block [13, 14]. Thus, for each layer $l \in 1, \dots L$, the model calculates

$$\boldsymbol{z}'_l = \text{MSA}(\text{LN}(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1}; \qquad \boldsymbol{z}_l = \text{MLP}(\text{LN}(\boldsymbol{z}'_l)) + \boldsymbol{z}'_l. \tag{3}$$

Finally, to calculate the image representation, a LN is applied to the last class token: $y = \text{LN}(z_L^0)$.

One major benefit of this transformer architecture is that it has significantly less image-specific inductive bias than CNN models. This is because a convolutional layer is local and dependent on the neighborhood structure within the two-dimensional $H \times W$ space (though we have seen that with deep CNNs the overall receptive can be approximately global). In contrast, since MSA is fully connected, these self-attention layers are fully global.

4.2.3 ResNet

In addition to ViT, we use a variety of convolutional methods. The first is ResNet, which presents a residual learning framework to allow deeper neural networks [15]. The main challenge for deep neural networks in computer vision prior to the introduction of ResNet was not overfitting but instead that they were difficult to successfully optimize. To address this, ResNet introduces residual “shortcut" connections, as illustrated in Figure A2.

Formally, these connections allow the model to bypass the non-linear transformations by providing an identity function, that is, on the l-th layer, we have $\boldsymbol{x}_l = F_l(\boldsymbol{x}_{l-1}) + \boldsymbol{x}_{l-1}$. If we consider H_l to be the underlying mapping from $\boldsymbol{x}_l$ to $\boldsymbol{x}_{l-1}$, then we are effectively trying to learn the residual, i.e. $F_l(\boldsymbol{x}_{x-1}) = H_l(\boldsymbol{x}_{l-1}) - \boldsymbol{x}_{x-1}$. This is hypothesized to address the issue as solvers often have issues with many consecutive nonlinear layers; see [15] for an in depth discussion of the architecture.

4.2.4 DenseNet

We also use a DenseNet, a model whose signature contribution is that each layer in the network is not only connected to the one before it but to all previous layers [16]. As a result, all of the feature maps from the previous layer are used as in inputs for a given layer. Specifically, the l-th layer in the network is defined as $\boldsymbol{x}_l = H_l([\boldsymbol{x}_0, \dots, \boldsymbol{x}_{l-1}])$, where H_l represents a composition of operations. This dense connectivity has led to the model being referred to as a DenseNet. In their implementation, the authors define H_l as the sequence: batch normalization (BN), ReLU, and a 3×3 convolution. Placed in between these dense blocks are transition layers which correspond to the composition of BN, a 1×1 convolution, and a 2×2 average pooling layer. Generally, these transition layers are also used to downsample the image with a compression factor. This architecture addresses the vanishing-gradient problem common in deep neural networks and also "strengthens feature propagation" [16].

4.3 Evaluation Method

For classification, we convert our target into an ordinal variable by binning it into C classes. This allows us to use multi-class cross entropy loss with mean reduction defined as follows:

$$\begin{aligned} l(x,y) &= \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N} w_{y_n}} l_n, \\ l_n &= -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})}. \end{aligned} \tag{4}$$

Here, x is the input, y is the target, w is the weight, N represents the batch size and C the number of classes. We define class weights w_c according to the inverse distribution of classes in our training sample. This ensures meaningful predictions, particularly given the unbalanced nature of our labels.

Although our focus was on the classification task, we did some initial exploration of the regression problem by modifying the output layers of our vision models and feeding the MOSAIKS embeddings into a linear regression. We utilize a mean-squared error loss (squared L2 norm) defined as follows:

$$l(x,y) = \frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^2. \tag{5}$$

Here, x is the input, y the target, and N the batch size.

5 Experiments

5.1 Experimental Details

We run our experiments using the Google Cloud setup provided by the CS 271 course, which means an n1-standard-8 instance (8 vCPUs and 30GB RAM running one NVIDIA T4 GPU). We developed our own Github repository and made use of `pytorch` and Huggingface model implementations. We trained all models with a batch size of 64 and fine-tuned for 20 epochs, saving the model with the best balanced validation accuracy. The ResNet and DenseNet models were pretrained on ImageNet-1k and the ViT was pretrained on ImageNet-21k. We used `AdamW` to set the learning rate with a initial values of `lr=1e-4` and `eps=1e-8`. The learning rate was reduced after hyperparameter tuning to enable more stable updates. We were concerned about overfitting in our image models, and thus chose `AdamW` for our optimizer as it uses weight decay.

We split the images into train and validation sets according to a 85/15 percent split. We specify a cross entropy loss using class weights according to the inverse class frequencies in order to ensure a balanced objective. In addition, for each image we specify a series of transforms. For DenseNet and Resnet, we each image to a PIL image, resize to a 224x224x3 image, and normalize the RGB values to have mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] (these are the values observed in ImageNet). We introduce some stochasticity during training by using a randomized crop and adding a randomized horizontal flip. For logistic regression and our basic convolutional network, we apply the pretrained ViT feature extractor for both train and validation, which we found to perform better than the transforms above, though we omit these results for brevity. Finally, for ViT, we also use its pretrained feature extractor. An example of an image alongisde its feature extracted image is presented in A4.

We also design a joint fusion model, enhancing our vision models with image-specific metadata, which is concatenated with the output of the vision model and fed into a multi-layer perceptron. In our implementation, we include either a 64-dimensional embedding layer of an image's country ("Country Embedding") or the 4000-dimensional MOSAIKS generalized features ("MOSAIKS").

5.2 Quantitative Results

We report the balanced validation accuracy for our experiments in Table 1. Our top-performing model across all specifications is some version of the pretrained Vision Transformer. For both quintile binnings, the plain ViT performs best without metadata augmentation; in contrast, in the two remaining binnings, the ViT with country embeddings performs best. Notably, the ViT enhanced with MOSAIKS always underperforms the best ViT model, though its average performance across all specifications is consistent with the other ViT formulations. For the BMI task, ViT performs twice as well as random chance and significantly outperforms all of our baseline models. Similarly, ViT performs 50 percent better than random chance for the under 5 mortality rate task. Importantly, the accuracy levels underscore the intrinsic difficulty of the task and suggest space for further research to refine and improve on these results.

Table 1: Balanced validation accuracy (%) by model and outcome binning

		Body Mass Index		Under 5 mortality rate	
Model	Metadata-enhanced	Quintile	CDC	Quintile	Severity
Baselines:					
Random		20	33.3	20	33.3
MOSAIKS		25.7	46.8	24.9	38.8
Logistic Regression		30.1	54.7	24.9	42.6
Two-Layer CNN		36.7	50.1	24.3	39.0
Two-Layer CNN	Country embedding	33.4	56.7	27.4	42.3
Vision Models:					
ResNet		39.2	58.8	30.4	45.1
DenseNet		**40.8**	59.2	29.7	47.0
Vision Transformer		**40.8**	61.5	**32.9**	49.6
Vision Transformer	Country embedding	39.5	**66.7**	30.6	**50.3**
Vision Transformer	MOSAIKS	39.8	64.1	32.3	49.8

One important acknowledgement is that ViT was pretrained on the most extensive and robust image dataset (ImageNet-21k) relative to the other pretrained models and our baselines were not pretrained. Due to computational limitations, we are not able to pretrain these models on ImageNet-21k (which consists of greater than 14 million images and over 21,000 classes). Moreover, a model architecture such as ViT is simply too large to train exclusively on our main sample. Nonetheless, ViT is also the most state-of-the-art and complex model that we use, and it is likely mucch better situated to successful pretraining on ImageNet-21k than our baselines or the other vision models. Thus, we do not believe that this is a significant limitation to the interpretation of our findings.

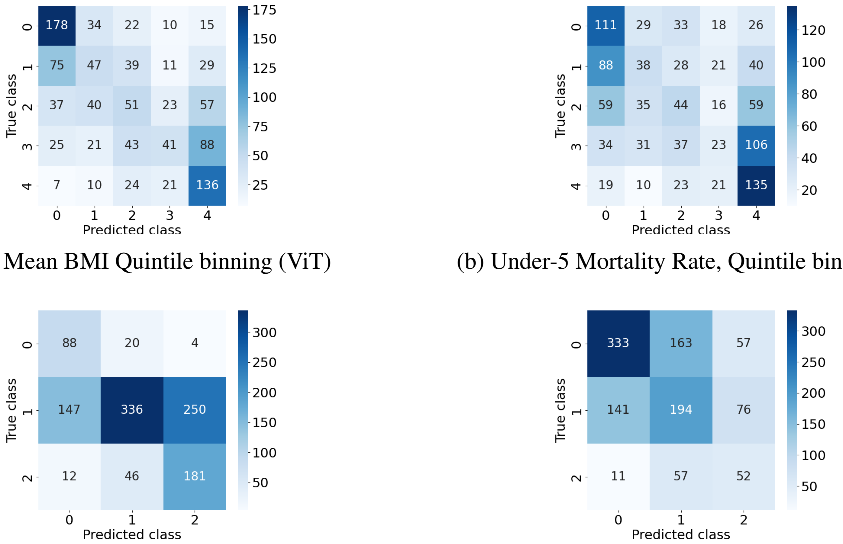
In Figure 5, we report confusion matrices for the top-performing model for each of our prediction tasks. This enables us to calculate the specificity and sensitivity for each class, where

$$\text{specificity}_c = \frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c} = \text{TPR}_c, \qquad \text{sensitivity}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} = \text{TNR}_c, \tag{6}$$

which we report in Table 2. Notably, the BMI models appear most successful at the extremes: for both specifications, the specificity is highest at the bottom- and top-most class. Indeed, the models appear to have the most difficulty disentangling the images from the middle of the distribution (as seen by specificities around 0.2 for the interior classes in the quintile approach). This is an important result, as from a public health perspective we are most interested in identifying regions that are malnourished rather than obtaining the most accurate functional distribution of weight over the images. A similar dynamic exists for the UMF rate quintile, with extremely low interior specificities and substantially higher values at the extremes. This is particular noteworthy as detecting regions with high UFM rates is incredibly important. Unfortunately, this result is not replicated in the UMF rate severity model, with the specificity actually tapering off for the higher class.

Table 2: Class-specific specificity and sensitivity for top-performing model (ascending order)

Figure	Outcome	Model	Specificity	Sensitivity
5a	BMI quintile	ViT	$[0.69, 0.23, 0.25, 0.19, 0.69]$	$[0.83, 0.88, 0.85, 0.92, 0.79]$
5b	UMF rate quintile	ViT	$[0.51, 0.18, 0.21, 0.1, 0.65]$	$[0.77, 0.88, 0.86, 0.91, 0.74]$
5c	BMI CDC	ViT w/ country	$[0.79, 0.46, 0.76]$	$[0.84, 0.81, 0.7]$
5d	UMF rate severity	ViT w/ country	$[0.6, 0.47, 0.43]$	$[0.71, 0.67, 0.86]$



(a) Mean BMI Quintile binning (ViT)

(b) Under-5 Mortality Rate, Quintile binning (ViT)

(c) Mean BMI CDC binning (ViT with country embeddings)

(d) UMF rate severity binning (ViT with country embeddings)

Figure 5: Confusion matrices for top-performing models

5.3 Qualitative Results

In order to elucidate what aspects of the images were important to the classification decisions, we saved the attention matrices within the ViT for each image. Intuitively, these capture what specific areas of the image the attention heads learned were optimal to focus on during the prediction task. As there are multiple attention heads and layers, we average attention across the heads and then recursively multiply the (normalized) attention weight matrices together and cast the resulting matrix into the original image space.

Figure 6 provides an example of this calculation with the original image on the left, the attention mask on the right, and attention map in the middle (calculated as the element-wise multiplication of the two). Interestingly, the main activation in the attention matrix is in the bottom right corner, where there are two villages. In addition, there appears to be a vertical line in the attention mask on the left side of the screen, tracing the road network. In contrast, the least relevant portions of the image according to the attention matrices appear to correspond to empty landscape. This is encouraging, as it is in accord with our initial hypotheses about the potential signal contained in road networks and village structure. We provide additional examples of visualized attention in the appendix (see A5). These examples further illustrate the model attending to river and road networks, the presence of a lake and settlement surrounding it, and outline of the coastline.

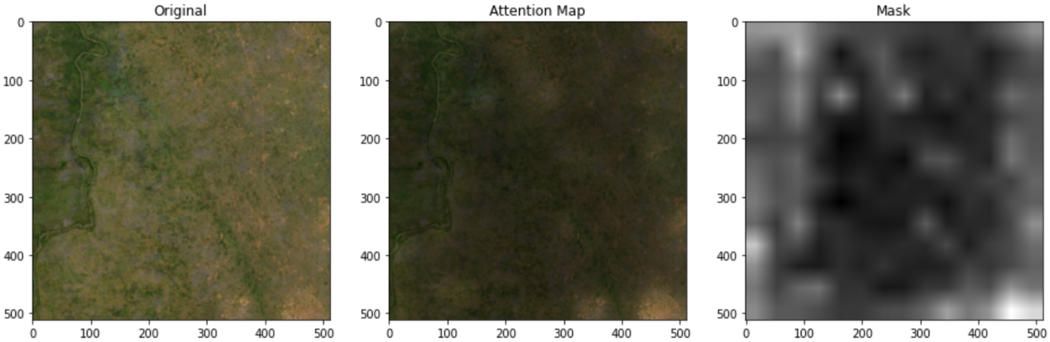


Figure 6: Example of visualized self-attention from ViT BMI model

5.4 Regression Specification

We conducted an initial exploration of the regression problem using the MOSAIKS embeddings fed into a linear regression as well as each of our vision models. We encountered significant difficulty with this approach; our models struggled to learn and yielded predictions with very little separation. In particular, the predictions were almost always extremely tightly clustered around the mean. Notably, L2 loss is a much more unstable and more difficult quantity to optimize than cross entropy loss, which was used for the classification. This makes intuitive sense: the precise values of the predictions do not matter in the classification, so long as the correct class receives the highest score. In contrast, the magnitude of the score is the very quantity we are trying to accurately predict in L2 loss. This makes the L2 loss formulated in the regression problem significantly less robust. Consequently, it seems likely that we have insufficient data and computational power to properly learn the continuous mapping from images to our targets.

6 Conclusion

We use satellite imagery to predict mean BMI and child mortatily rate in regions across western Africa. We define and make use of multiple models, including Vision Transformers, ResNet and DenseNet, as well as the MOSAIKS featurization model. We make mean BMI predictions for both a quintile split as well as the CDC-designated BMI split; we make the infant mortality predictions for a quintile split and a more intuitive, ternary split based on severity. Our best model on the CDC-designated BMI split achieves an accuracy of 66.7%, doubling the accuracy of a random baseline; our best model on the severity binning for the infant mortality rates achieves an accuracy of 50.3%, a 51% improvement over the random baseline.

There are many future avenues of research. First, future work should look into further model interpretability and visualization techniques in order to better understand what image features are most valuable for these predictions. Moreover, the metadata augmentation could be further expanded to include longitude and latitude coordinates as well as historical economic and political indicators. Work on predicting other health indicators such as maternal mortality rate could also be pursued.

Finally, with access to more data and computational power, it would be interesting to evaluate the model performance using a broader set of countries from a roughly equivalent time period. All of this work would advance the goal of eventually being able to predict regions that require additional healthcare resources and assistance without much reliance on costly and limited survey methods.

References

[1] Ryan Engstrom, Jonathan Samuel Hersh, and David Locke Newhouse. Poverty from space: using high-resolution satellite imagery for estimating economic well-being. Technical Report 8284, World Bank Policy Research Working Paper, 2017.

[2] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[3] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

[4] Luna Yue Huang, Solomon M Hsiang, and Marco Gonzalez-Navarro. Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. Technical report, National Bureau of Economic Research, 2021.

[5] Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 616–625, 2018.

[6] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*, 2017.

[7] Jean-Emmanuel Bibault, Maxime Bassenne, Hongyi Ren, and Lei Xing. Deep learning prediction of cancer prevalence from satellite imagery. *Cancers*, 12(12):3844, 2020.

[8] Joshua J Levy, Rebecca M Lebeaux, Anne G Hoen, Brock C Christensen, Louis J Vaickus, and Todd A MacKenzie. Using satellite images and deep learning to identify associations between county-level mortality and residential neighborhood features proximal to schools: A cross-sectional study. *Frontiers in public health*, 9, 2021.

[9] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12, 2021.

[10] Adyasha Maharana and Elaine Okanyene Nsoesie. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA network open*, 1(4):e181535–e181535, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[13] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[14] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

A Appendix

A.1 Acknowledgements

Special thanks to our mentors, Professor Pascal Geldsetzer and Ali Lenjani.

A.2 Figures and Tables

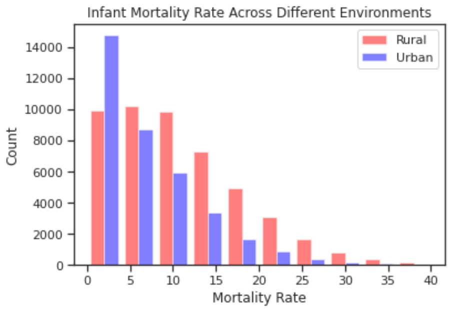


Figure A1: Variance of child mortality rates across differing environments

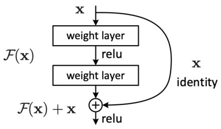


Figure A2: An example of a residual connection [15]

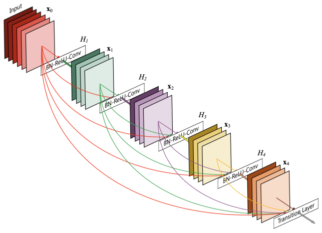
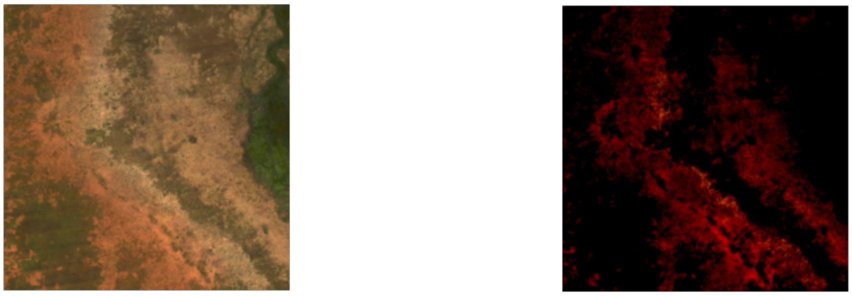


Figure A3: DenseNet architecture [16]



(a) An example input image before feature extraction

(b) Same image after feature extraction

Figure A4: ViT feature extraction of an input image. This feature extration applies a variety of transforms to systematically process an image for the model. This includes normalizing the mean and standard deviation to match the pretrained data (ImageNet-21k). In order to plot the image, values less than 0 and greater than 1 are clipped.

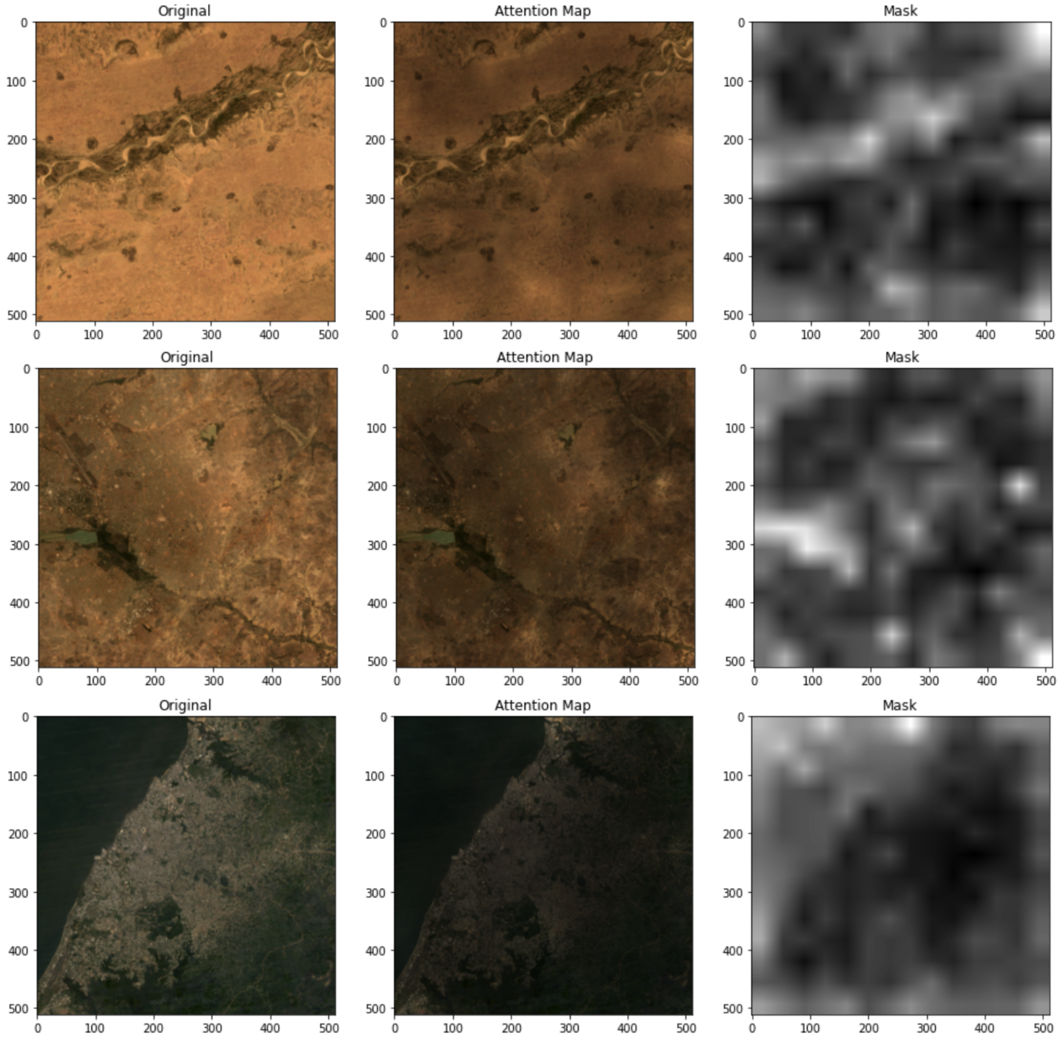


Figure A5: Additional examples of visualized attention matrices. The first row shows the attention paid to a river and road network, the second to the presence of water and a settlement nearby, and the third appears to take note of the coastline.

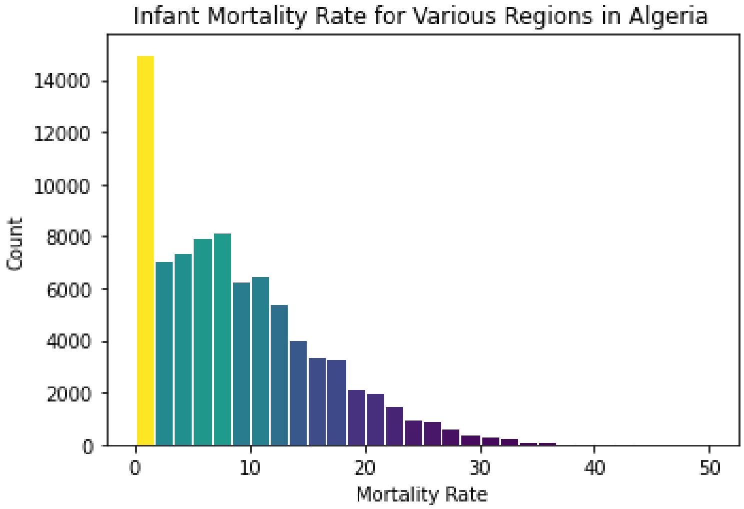


Figure A6: Distribution of infant mortality rates across Algeria

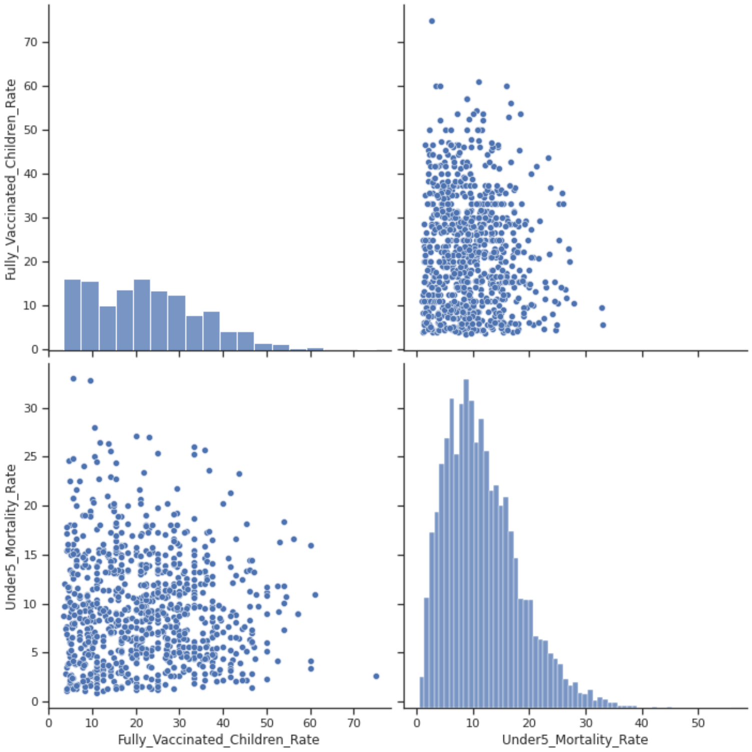


Figure A7: Vaccination vs. under 5 mortality rate