Predicting Micronutrient Deficiency with Publicly Available Satellite Data

Elizabeth Bondi-Kelly,[1] Haipeng Chen, [2] Christopher D. Golden, [3] Nikhil Behari, [4] Milind Tambe [5]

[1] MIT / University of Michigan, ecbk@umich.edu
[2] William & Mary
[3] Department of Nutrition, Harvard T.H. Chan School of Public Health
[4] MIT Media Lab
[5] Center for Research on Computation and Society, Harvard University

Abstract

Micronutrient deficiency (MND), which is a form of malnutrition that can have serious health consequences, is difficult to diagnose in early stages without blood draws, which are expensive and time-consuming to collect and process. It is even more difficult at a public health scale seeking to identify *regions* at higher risk of MND. To provide data more widely and frequently, we propose an accurate, scalable, low-cost, and interpretable regional-level MND prediction system. Specifically, our work is the first to use satellite data, such as forest cover, weather, and presence of water, to predict deficiency of micronutrients such as iron, Vitamin B12, and Vitamin A, directly from their biomarkers. We use real-world, ground truth biomarker data collected from four different regions across Madagascar for training, and demonstrate that satellite data are viable for predicting regional-level MND, surprisingly exceeding the performance of baseline predictions based only on survey responses. Our method could be broadly applied to other countries where satellite data are available, and potentially create high societal impact if these predictions are used by policy makers, public health officials, or healthcare providers.

Introduction

More than 2 billion people worldwide, including 340 million children (Keeley, Little, and Zuehlke 2019), are affected by micronutrient deficiencies, or the lack of vitamins and minerals required by the body for healthy functioning and development (Micha et al. 2020). These micronutrient deficiencies, hereafter referred to as MND, further drive the global burden of disease but remain difficult to diagnose since the effects often become visible only when the deficiency is severe (von Grebmer et al. 2014). From a public health perspective seeking to reduce MND prevalence throughout a population, it is important to identify regions at risk of MND. However, due to the difficulty of diagnosing MND, regions with MND are unclear to public health organizations until direct measurements are made, such as blood draws to measure biomarkers and/or surveys/questionnaires. Unfortunately, these blood draws and surveys are costly and time-consuming, and furthermore, quantifying micronutrient levels in a blood sample requires limited, specialized laboratory equipment, leading to infrequent data collection.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Due to the difficulty in both types of data collection, we seek a new data source that may be more scalable, such as satellite data (i.e., data products derived from raw satellite imagery). This may at first seem unrelated, as MND status is unique to an individual, pertaining to an individuals' nutrition, disease status, and other characteristics which cannot be viewed by satellite. Indeed, prior work applying artificial intelligence (AI) techniques to satellite data, e.g., in estimating crop type (Gadiraju et al. 2020), often search for features directly observable by satellite. Predicting an indirect feature such as MND prevalence brings additional technical challenges, including choosing relevant satellite data, linking a limited amount of ground truth data from individuals to satellite data to train machine learning models, and supporting interpretability for public health experts.

Contributions: Through our novel system, we establish that satellite data can be used to predict MND at a regional level despite these challenges. In fact, our system is the first to predict MND from a regional level, as measured directly from real-world, ground truth biomarkers, using satellite data. This involves i) aggregating individuals' MND states from biomarker data over geographic regions to align with satellite data, ii) using segmentation to generate custom features of importance, specifically market locations in this case, iii) providing scalablity with automatic feature selection, which performs comparably to expert feature selection, and iv) two prediction paradigms to handle the challenges that arise from limited ground truth data: logistic regression, which also naturally handles the pressing need for interpretability of predictions in the field, and multi-layer perceptron with domain adaptation. Not only does this system achieve good accuracy, but this also results in improved performance compared to the baseline of survey-based predictions. We believe this MND detection system could be broadly applied to other countries where satellite data are available, potentially leading to more information for public health interventions and high societal impact.

Background and Related Work

AI for Social Impact and Satellite Data: Existing applications of AI related to nutrition include food security, agriculture, food rescues, and even foodborne illnesses (Shi, Wang, and Fang 2020). Some of this literature relies on

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/aaai.12080

This article is protected by copyright. All rights reserved.

Author Manuscript

satellite and other remotely-sensed images, such as agricultural productivity assessments and planning (Nakalembe 2020). Land cover mapping (Poortinga et al. 2019) and socioeconomic status prediction (Ayush et al. 2020) have also been explored. However, these factors are arguably directly visible in satellite data, e.g., to predict socioeconomic status, Ayush et al. (2020) search for objects directly in satellite data, such as trucks. Dengue fever prediction in Abdur Rehman, Saif, and Chunara (2019) is based on identifying features such as standing water locations (mosquito habitat) and roads (human presence). While dengue status is not directly visible, these direct causes are. MND prediction is less direct, as it may depend on disease *and* nearby agriculture, forests, etc.

Possible Causes of MND: The causal mechanisms of MND are complex, but there are multiple factors that likely influence MND, including environmental (e.g., forest presence), epidemiological (e.g., malaria), and socio-economic factors. One of the primary environmental factors studied for its impacts on MND is forests. Generally, research indicates that access to forests may improve dietary diversity. Dietary diversity is an assessment of the range of food groups consumed over a period of time that is typically used as a proxy for sufficient nutrient intake (Steyn et al. 2006), which *is typically measured using survey responses detailing foods consumed.* Forests may directly support dietary diversity, e.g., from bushmeat and wild fruits, provide an additional source of income, e.g., through the sale of forest products, or support crop and livestock production (Sunderland, O'Connor et al. 2020). A study on children's diets across 27 developing countries, including Madagascar, finds that close proximity to forests improved the household prevalence of Vitamin A- and iron-rich foods by 11% and 16%, respectively (Rasolofoson et al. 2018). Ickowitz et al. (2014), one of the most similar studies to ours, analyze dietary diversity, fruit and vegetable consumption, and animal source food consumption in children using satellite data such as tree cover, road location, climate, and urban population information.

As an example of socioeconomic factors, Koppmair, Kassie, and Qaim (2017) show that access (as measured by distance) to food markets in Malawi plays an important role in supporting dietary diversity, particularly for farm households. Markets may directly provide food, and/or may provide additional sources of income for local residents through agricultural and livestock production sales, which can indirectly improve dietary diversity. Agriculture, livestock, and water supply also play an important role in health and nutrition (Brown et al. 2014).

While these methods imply that satellite data can contribute towards predicting MND, dietary diversity depends only on foods consumed, which may be directly observable from satellite imagery (e.g., crops or forests). Biomarkers may involve further subtleties, such as individual characteristics or disease. We use additional features as a result.

Data Description

Ground Truth Data: Ground truth data were collected by Golden et al. (2020) in 2017-2018 in four distinct ecological regions in Madagascar, denoted as the Central Plateau (CP), Southwest (SW), Southeast (SE), and West Coast (WCO) (see Fig. 1). CP is at a high elevation, SW is arid, SE is a mid-altitude rainforest, and WCO is seasonally dry.

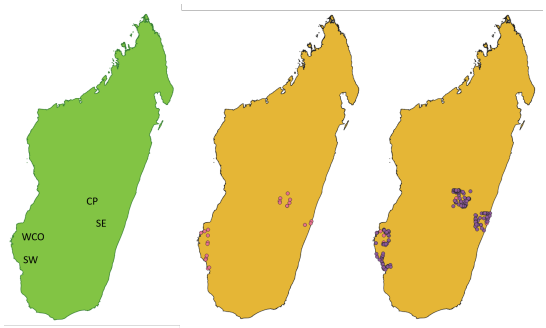


Figure 1: Regions studied in Madagascar (left), known (center) and predicted (right) markets in these regions.

In this paper, we will focus on the survey responses and biomarker data from blood samples that were collected in Golden et al. (2020). Surveys were provided to individuals in households, small groups, and more. In total, responses were collected from 6292 individuals from 1125 households within 24 communities in CP, SE, SW, and WCO. Biomarker levels from blood draws were also collected from a subset of these individuals. We denote the set of individuals by $p \in \{0, 1, ..., P\}$. Each individual has an underlying MND state, d_p, based on a biomarker level, m, that is thresholded by t, derived from public health literature. In Table 1, we include the thresholds used to define MND in this paper, though we are unable to provide raw data publicly. Therefore, individual p has $d_p = 1$ if $m < t$ and 0 otherwise. After combining data from blood draws with surveys and household GPS locations, we have 2458 samples.

During this data collection process, Golden et al. (2020) followed all procedures to minimize the risk to local populations involved as subjects in the study, as detailed in our approved IRB protocol from the Harvard T.H. Chan School of Public Health (IRB16-0166). This included gaining informed consent for all study-related protocols, including the future cross-referencing of biological data with remotely sensed data products to improve the targeting of public health responses. To briefly summarize this process, a community meeting was held to explain the study using speeches. The research team then visited sampled households to invite individuals to participate. The prospective participants were provided more information if they expressed interest. Furthermore, data are de-identified to limit the risk of breaches of confidentiality, and we follow Harvard IRB protocols to further minimize risk. Gaining informed consent does not automatically alleviate concern of data misuse and inadvertent consequences; nevertheless, we took all necessary precautions to protect human subjects in the study. Please see Golden et al. (2020) for further details.

Satellite Data: Based on the causes of MND in Related Work, we select publicly available satellite data, much of

Author Manuscript

This article is protected by copyright. All rights reserved.

MND	Biomarker	Values
Iron [a]	Ferritin	< 30 ng/mL
Vitamin A [b]	Retinol	< 0.20055 mg/L
Vitamin B12 [c]	B12	< 300 pmol/L

Table 1: Micronutrient deficiency thresholds.

[a] https://www.who.int/vmnis/indicators/serum_ferritin.pdf
[b] https://apps.who.int/iris/bitstream/handle/10665/44110/9789241598019_eng.pdf
[c] USDA

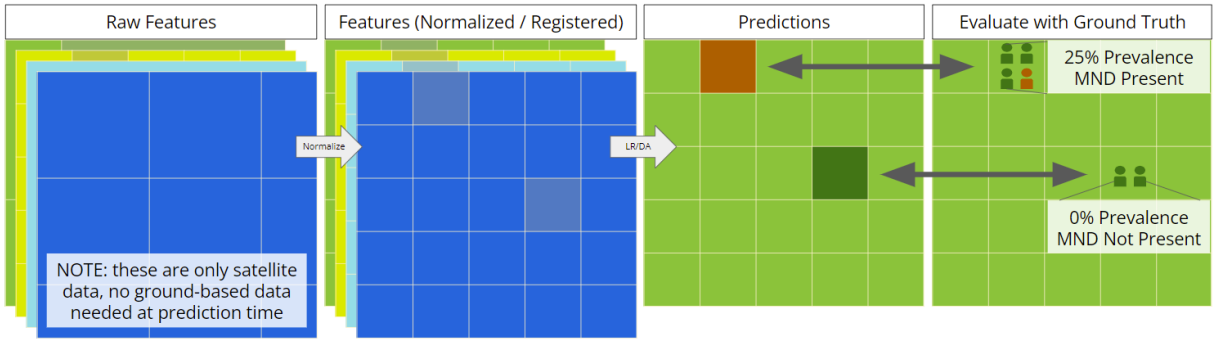


Figure 2: Illustration of using satellite data, which is first normalized and registered, as features to predict MND. Compare to pixel-level labels derived from individual MND statuses. In this illustration, both predictions are correct.

Feature	Collection Time	Google EE
Livestock Population Density (Robinson et al. 2014)	2010	
Crop Cover (Xiong et al. 2017)	2015	
Elevation (NASA JPL 2020)	2000	✓
Fire (Giglio and LANCE FIRMS 2016)	2016	✓
Fishing Hours (Kroodsma et al. 2018)	2016	✓
Forest Cover (Shimada et al. 2014)	2017	✓
Forest Change (Hansen et al. 2013)	2017	✓
Landcover (Buchhorn et al. 2020)	2017	✓
Nighttime Lights (Elvidge et al. 2017)	2017	✓
Population Density (CIESIN 2017)	2015	✓
Presence of Water (Pekel et al. 2016)	1984-2019	✓
Weather (McNally et al. 2017; NASA GSFC HSL 2018)	2017	✓
Crop Production (International Food Policy Research Institute 2020)	2017	
Markets (Golden et al. 2020)	2017-2018	
Healthcare Sites (Humanitarian Data Exchange 2020)	2020	

Table 2: Satellite data sources, collection time, and availability on Google Earth Engine.

Author Manuscript

This article is protected by copyright. All rights reserved.

which is derived from raw satellite imagery, e.g., using machine learning. In particular, we used the data shown in Table 2. We acknowledge the use of data and/or imagery from NASA's Fire Information for Resource Management System (FIRMS) (https://earthdata.nasa.gov/firms), part of NASA's Earth Observing System Data and Information System (EOSDIS).

Once we collect these features (in the form of images) at the sites of clinical data collection, we resample the images to a uniform resolution of about 25x25 m for one pixel, at a size of 308x308 pixels. This provides us with 23 images total with 86 features each (as image bands). After collecting all satellite data, we normalize each feature to within [0, 1], regardless of whether it was binary, categorical, or continuous. We then do imputation by taking the nearest neighbor if there are any missing data in the feature.

Problem Description and Aggregation

Given the values from satellite data for a pixel as *input*, our goal is to predict MND presence (classification) or prevalence (regression) in that pixel as the *output*. Ground truth *labels* are derived from biomarkers in blood samples.

Define Grid with Satellite Data: More specifically, we represent the input, i.e., the satellite data, via a multidimensional image array, S. There are 23 S in our dataset, as the ecological regions are large. Therefore, we add an overall image index, $S^{l,r}$, where r represents the current region, and l represents the image index within that region. Each $S^{l,r}$ is indexed by i for rows (y-axis), j for columns (x-axis), and k (z-axis) for features, i.e., the individual satellite data features such as forest cover, weather, and presence of water.

Aggregation to Link Data: To link the two data sources, we rely on locations. Each p (individual, see Data Description) is associated with some g_p, a geographic coordinate. Each $S^{l,r}_{i,j}$ is associated with a set of geographic coordinates, $G^{l,r}_{i,j}$. We may now find the set of individuals, $P^{l,r}_{i,j}$, whose locations fall within each pixel, such that $g_p \in G^{l,r}_{i,j}$. We find their underlying MND states, d_p, to calculate MND prevalence, the percentage of individuals who have MND as defined by biomarker levels. This prevalence, $v^{l,r}_{i,j}$, is our label:

$$v^{l,r}_{i,j} = \frac{\sum_{p \in P^{l,r}_{i,j}} d_p}{|P^{l,r}_{i,j}|}, \tag{1}$$

where $|P^{l,r}_{i,j}| = \sum_{p \in P^{l,r}_{i,j}} 1$ is the cardinality of set $P^{l,r}_{i,j}$. We may threshold $v^{l,r}_{i,j}$ for a classification task, or predict the explicit value directly as a regression task. Please see Fig. 2 for an illustration. In our dataset, this leads to 300-500 pixel labels, which is only about 0.02% of pixels.

Formally, our goal is to train a *region-specific* ML model $f^r_\omega(\cdot)$ parameterized by ω for each of the 4 ecological regions, where given input training data $S^{l,r}_{i,j}$ in the training set, the model is optimized to minimize the discrepancy between prediction $\hat{v}^{l,r}_{i,j} = f^r_\omega(S^{l,r}_{i,j})$ (see Fig. 2) and the *ground truth label* $v^{l,r}_{i,j}$: $\min_\omega \mathbb{E}_{S^{l,r}_{i,j} \in S^r_{tr}} D(\hat{v}^{l,r}_{i,j}, v^{l,r}_{i,j})$ where $D(\hat{v}^{l,r}_{i,j}, v^{l,r}_{i,j})$ could be, e.g., mean squared error (MSE) for regression, or cross-entropy (CE) for classification. $\mathbb{E}_{S^{l,r}_{i,j} \in S^r_{tr}}$ is an expectation taken over all pixels in the training set S^r_{tr} in region r, for each micronutrient. We assume the data are *i.i.d.*

Prediction Methodology

Market Detection: As discussed in Related Work, the presence of markets is an important factor for MND. We would consequently like to add markets as an extra feature on top of the existing satellite data products. Yet, it is difficult to know where all markets are located in Madagascar. We only know of those specifically mentioned during the focus group surveys conducted in Golden et al. (2020).

To add this, we therefore start by comparing the known market locations from the survey data responses with satellite data, and infer that the number of buildings within town clusters and the proximity to roads may be used as predictors of market presence in Madagascar. Specifically, we determine empirically that 20 buildings and one road within about 0.8 km^2 are highly indicative of market presence.

In order to apply these thresholds in an automatic market detection pipeline, we first have to locate roads and buildings. While OpenStreetMap (OSM)[1] provides building and road segmentation data, it is not always complete. This is especially true in our regions of interest. As a result, we train a satellite image-based segmentation model.

For ground truth data to train this segmentation model, we use nearby OSM building labels *where they are more complete*. In particular, for each of the four regions in Madagascar, we automatically identify the closest densely-clustered OSM building labels to the known market locations. These labels are saved to the building segmentation training set, along with high-resolution images from the Google Maps Static API[2]. For each region, the training dataset contains roughly 100-200 training images and at least 500 corresponding OSM building labels across all images. Each individual image has 600x600 pixels, with a 0.46 m resolution.

For the building segmentation model, we use a U-Net convolutional network (Ronneberger, Fischer, and Brox 2015) with a ResNet-34 encoder pretrained on ImageNet. The U-Net architecture, originally developed for biomedical image segmentation, is commonly used for satellite image segmentation, and is particularly useful for training on smaller training sets such as the sparse OSM building label data. The satellite image training set is augmented with random flips, rotations, and resizes. Binary cross entropy is used as the loss function, and we use the Adam optimizer with a learning rate of 1e-2. The model is trained using a batch size of 16. Results are shown in Fig. 1. The building segmentation model and thresholding achieves 0.86 precision in detecting the ground-truth markets from survey data. We include these as features in our data by drawing radii of multiple distances around each market, so that pixels in this layer represent the number of markets within a certain radius. We create these radii masks given healthcare center coordinates

[1] www.openstreetmap.org

[2] developers.google.com/maps/documentation/maps-static

Author Manuscript

This article is protected by copyright. All rights reserved.

(Humanitarian Data Exchange 2020) as well, bringing us to 90 total features. While we focus on markets here, *this segmentation process could be applied to generate other satellite image-based features that do not already exist*, such as custom landcover maps.

K-Medoids-based Feature Selection: It is helpful to have many features, but not all features are necessarily informative. The risk of overfitting when using all 90 features can be large when dealing with limited data. A straightforward idea is to use knowledge from domain experts to select only features that are most important for predicting MND in a particular region. However, this introduces two more issues. First, the feature importance of different regions may vary drastically due to different ecologies. In Madagascar, for example, certain agriculture, such as pulses, are only present and predictive of MND in some regions. It would require a significant amount of manual work to specify the set of important features for each area. Second, the causal mechanisms behind MND are not fully understood. Therefore, it is critical to come up with an automatic feature selection procedure that effectively filters out uninformative features with minimal manual effort.

We start by removing any features that are always 0 throughout the full dataset (i.e., $S_{i,j,k} = 0, \forall i, j$), leading to 69 features. We then use the K-medoids clustering method (Park and Jun 2009) to group highly correlated features. Each point in our space is a vector of individual pixel values in an image (representing a feature), such that the dimension of the space is the number of pixels. We use Pearson's correlation coefficient as the distance metric between features. Similar to K-means clustering, K-medoids clustering also aims at partitioning the data points (i.e., features) into different clusters. Both minimize the sum of distances between points labeled to be in the same cluster and a point designated to be the center of that cluster. However, K-means uses the central position (centroids) as the designated point, while K-medoids uses a point that actually exists in the set of data points (i.e., an existing satellite data feature). As such, we are able to use the medoid feature to represent the group of correlated features, preserving interpretability.

We post-process the image data, selecting the 300-500 (0.02%) ground truth pixels to form a feature matrix.

Prediction with Logistic Regression: We first use a simple but effective logistic regression model. We choose logistic regression as one of the underlying ML models in this paper, due to its following advantages. First, it has fewer weights compared to other models such as deep neural networks, and therefore is less prone to overfitting. This is particularly important given the limited amount of data we have and the high-dimensional feature space. Second, it is interpretable by itself (as shown in experiments, e.g., Fig. 6), where the weights ω of different features directly indicate the importance of the features in determining the prediction outcome. Moreover, compared to post-hoc model-free explanation methods such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), which only provide *instance*-level explanations, the weights of logistic regression models imply feature importance at an *aggregated* level, which we show could provide important insights to public health experts. We primarily focus on region-specific prediction for tailored interpretation and results, but we also train using all regions' training data combined and predict on each regions' test set, which we call Naively Combined.

Prediction with Multi-layer Perceptron and Domain Adaptation: Another strategy to address limited training data is domain adaptation (Huang et al. 2006), which allows us to use data from all 4 ecological regions as follows: The target domain is the region of Madagascar in which we are making our predictions. The source domains are the other 3 regions, which we would like to use for augmentation. We project all 4 into a domain-invariant latent representation with a single hidden layer (5 neurons) and the loss function:

$$l = \alpha * l_{src} + l_{tgt} + \lambda * l_{transfer} \quad (2)$$

where l_{src} and l_{tgt} are the binary cross-entropy loss in the source and target domains. $l_{transfer}$ is the CORAL loss (Sun, Feng, and Saenko 2016) between the source and target domains. α and λ are hyperparameters, and are tuned to be 0.1 and 0.01, respectively, out of $\{0.01, 0.1, 1, 10\}$. Finally, we predict on the target domain test set.

Results

We present experimental results using 4-fold cross-validation (i.e., data from one region are broken into 4 folds). Due to the limited amount of data, it is impractical to have more folds. We primarily report Area Under the Curve - Receiver Operating Characteristics (AUC-ROC, or AUC in short) to evaluate the MND classification tasks. Note that we only report the mean AUC values averaged over the 4 folds as the standard deviation becomes trivial for only 4 folds. All data collection and experimentation rely on the default, free resources on Google Colab[3], and training for all 4 folds takes less than 1 minute in general for both logistic regression and domain adaptation.

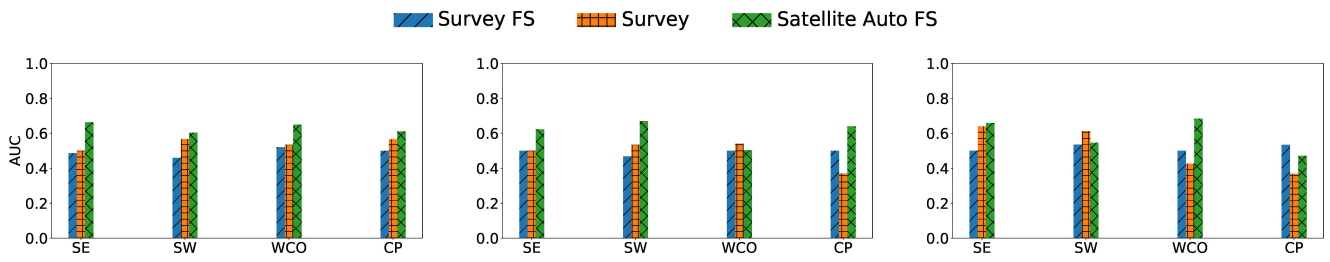
a) Is our prediction accurate? We compare with predictions made by survey data only, as is similar to prior work such as (Ickowitz et al. 2014). The results are shown in Fig. 3. For survey data, we tested two versions, the original, full amount of data, and a version with one simple level of feature selection. In this case, we selected features which we believed could reasonably be seen or inferred from satellite data. When comparing both survey-based predictions with our satellite data-based predictions, we can see that satellite data-based prediction is better in i) all 4 regions for iron, ii) 3 out of 4 regions for Vitamin B12, and iii) 2 out of 4 regions for Vitamin A. Where it does not outperform survey-based predictions, it performs comparably with significantly lower cost. Across all of the 4 regions and all of the 3 types of nutrients, the AUC value is higher than 0.6 in 10[4] cases, and is close to 0.5 for the other 2 cases. Meanwhile, the F1 scores of our predictions are on average 0.6 (ranging up to 0.9) and

[3]https://colab.research.google.com

[4]Please note that some of these statistics may slightly fluctuate, e.g., 9 instead of 10 cases sometimes.

Author Manuscript

This article is protected by copyright. All rights reserved.



(a) Iron deficiency (b) Vitamin B12 deficiency (c) Vitamin A deficiency

Figure 3: Comparison of survey-based (with or without feature selection) and satellite data-based MND prediction by regions.

Feature Description	Frequency
Chicken population density	9
Cattle population density	8
Net shortwave radiation flux	7
Presence of market within 7.5 km	6
Soil moisture in 100 - 200 cm underground	6
Soil temperature in 10 - 40 cm underground	5
Near surface wind speed	4
Surface pressure	4
Fire (temperature of pixel)	4
Presence of market within 3.75 km	3

Table 3: Frequency of each feature appearing in either the top 3 positive or negative coefficients. The 10 (out of 21) features with the highest appearance frequencies are shown.

are also comparable to those based on surveys. Recall is also important, as false negatives may lead to resources allocated away from people who truly have MND. Generally, recall is comparable to AUC for these data. However, it is higher in some cases. For example, for iron deficiency in region SE, recall is nearly 0.9.

Furthermore, satellite data-based regression results are comparable to survey-based regression. We also report the regression results in Fig. 4. We can see that the satellite imagery-based regression results are still comparable to the two versions of survey-based regression. In particular, MAE of our method ranges 0.16-0.19 in iron, 0.18-0.35 in Vitamin B12, and 0.19-0.29 in Vitamin A, which are reasonable considering the range of the regression task is $[0, 1]$ and the means are 0.21, 0.36, and 0.20, respectively. The AUC, F1-score, and MAE results all together demonstrate that our predictions are reasonably accurate as well.

b) Which features are important for MND prediction? As logistic regression is considered an inherently interpretable model, we focus our analysis on the weights of each variable, particularly those whose absolute values are largest. First, we build an "important features" list. For each region-specific model and each micronutrient (in total $3 \times 4 = 12$ cases), we record the features with the top 3 highest positive weights and negative weights. We aggregated statistics on the number of times that each feature appears in these "important features" lists in Table 3. From this, we observe that *market features are very important*, with market presence within 7.5 km with 6 appearances, and within 3.75 km with 3 appearances. We also observe other interesting trends, including that more forest fires are linked to greater rates of Vitamin A and B12 deficiency in the SE region (rainforest), but not in other regions that are less reliant on forest products, which may be a useful insight for public health experts. Fig. 6 illustrates this pattern for Vitamin A in SE.

Furthermore, we included multiple correlates to socioeconomic status, such as nighttime lights, i.e., images of Earth at night, where it is expected that highly populous and resourced areas have more light. We found that the correlation coefficient between nighttime lights data and ground truth iron deficiency is 0.127, implying that alone, it may not be highly correlated. The individual feature with the greatest correlation is the sugarcane crop, with 0.147. If we predict solely with sugarcane, we achieve an AUC of 0.428, which is less than our findings of about 0.6 for iron deficiency. This implies that we need the other factors as well in order to predict MND.

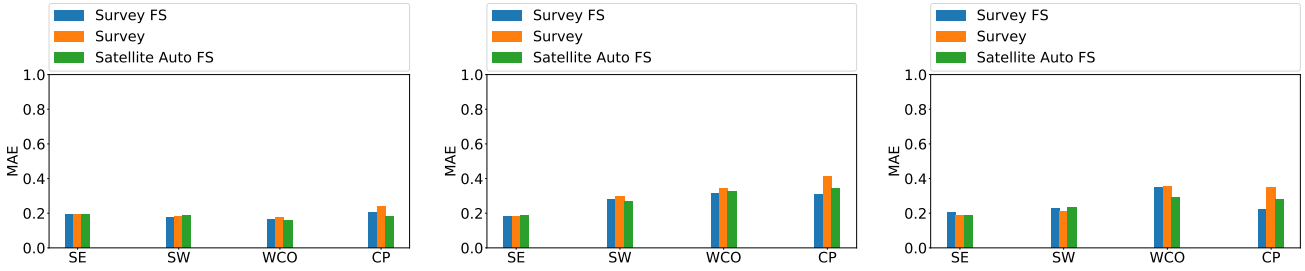
c) How does the automatic feature selection perform? To evaluate the performance of automatic feature selection (FS), we compare with two baselines. First, we consider the case where there is no feature selection apart from removing features which are completely zero (i.e., no data) (Satellite Remove 0 FS). We also compare with expert feature selection, in which a public health expert examines the features we propose, and groups them based on their knowledge[5]. They also select a representative feature for each of their groups (Satellite Expert FS). Finally, we consider the performance of our correlation and K-medoids-based algorithm (Satellite Auto FS). We show results for one of the regions (WCO) due to space limitation, but trends in other regions are similar. We can see that both Expert FS and Auto FS are better than the case where no FS is used, especially for Vitamin B12. In all three cases, Auto FS always performs comparably to Expert FS, as it does in other examples that are not included here, but Auto FS is more scalable.

We also compare the groups that are found by Auto FS and Expert FS. Very interestingly, we find that in the two

[5]Expert chose 21, which led us to select $K = 21$

Author Manuscript

This article is protected by copyright. All rights reserved.



(a) Iron deficiency (b) Vitamin B12 deficiency (c) Vitamin A deficiency

Figure 4: Regression results comparison between satellite imagery based and survey based predictions. All elements are the same as Fig. 3 except that y-axis now means MAE of the regression task. Note that in this figure, *lower bars imply better results*.

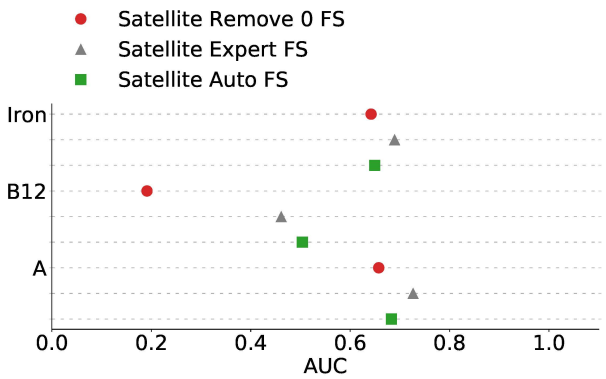


Figure 5: Comparison feature selection methods, including removing any features without data, human expert feature selection, and our K-medoids method, all in region WCO.

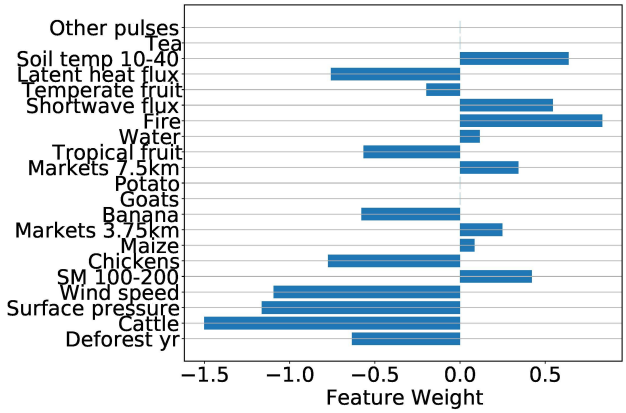


Figure 6: Logistic regression weights (x-axis) for Vitamin A, region SE. Positive numbers mean positive correlation with MND. Medoid feature names provided (SM: soil moisture).

methods, 8 out of 21 group centers overlap: banana, cattle, chicken, goat, maize, presence of markets within 7.5 km, surface pressure, and wind speed. This shows that our method is choosing features deemed important by a human expert as well. The above results well demonstrate that our proposed automatic feature selection method is an effective while scalable alternative to expert feature selection.

d) How do different prediction paradigms compare? We compare the region-specific logistic regression models (Satellite Auto FS), the logistic regression model version that combines training data from all of the regions (Naively Combined), and multi-layer perceptron with domain adaptation (Domain Adaptation). We present results from region CP. Here, and overall, we find that Vitamin B12 and Iron achieve better performance using Domain Adaptation, while Vitamin A achieves better performance using the logistic regression-based Satellite Auto FS or Naively Combined. This may be because each micronutrient differs slightly in its relevant factors, and factors may vary regionally (e.g., some regions are forested). Clearly, each method works well with limited amounts of data, but we acknowledge the tradeoffs in interpretability, and a potential lack of robustness in the model due to limited samples.

Conclusion and Discussion

In conclusion, satellite data are viable to use for MND prediction at a public health scale. We presented a system relying on the aggregation of individual MND states over geographic regions, a search for relevant features, such as markets, automatic feature selection, which performs comparably to human expert feature selection, and domain adaptation and logistic regression prediction models. This system worked well even with limited ground truth biomarker data.

Deployment Considerations While our system has not yet been deployed, we would like to emphasize several deployment considerations. This methodology would not replace surveys and blood samples collected among communities. Rather, we believe it should be used to cover gaps in that data collection, e.g., where data could not be collected, or in between collections. To do this, public health officials, policymakers, healthcare workers, or individuals can load publicly available, current satellite data and apply the existing model, without any survey or blood sample data. We can then update these models when another data collection occurs. This also applies for deployment in other countries. We plan to develop a web application to load satellite data at the desired time and location, and the current proposed

Author Manuscript

This article is protected by copyright. All rights reserved.

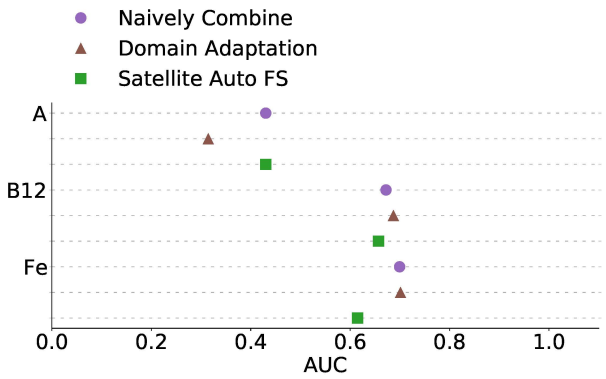


Figure 7: Comparing AUC of a logistic regression model trained by naively combining training data from all regions, a multi-layer perceptron with domain adaptation, and a region-specific logistic regression model, all in CP.

model, to provide predictions. We plan to iterate on this with potential users, including officials from Catholic Relief Services, Médecins Sans Frontières, and the Ministry of Health in Madagascar. In the meantime, code and satellite data are available[6], while ground truth data are withheld for privacy.

Future Work: We began preliminary experiments into sparse segmentation and spatial aggregation to further include spatial patterns in the prediction step, but they require further refinement before deployment. We also encourage the use of custom features, as illustrated with markets. Most importantly, we believe there is ample room for further research, and a great deal of promise for broad application to inform future public health interventions.

Acknowledgments

We are grateful for the support from the United States Agency for International Development (USAID) (Grant AID-FFP-A-14-00,008) implemented by Catholic Relief Services (CRS) in consortium with four local implementing partners in Madagascar; the Ren Che Foundation; the ARO under Grant Number: W911NF-18-1-0208; and the Harvard Center for Research on Computation and Society (CRCS). The views and opinions expressed in this paper are those of the authors and not necessarily the views and opinions of USAID, nor should be interpreted as representing the official policies, either expressed or implied, of ARO or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Thanks to Akshaya Annapragada for data imputation support. Conflict of Interest: The authors have no conflicts of interest to report.

[6]https://github.com/exb7900/mnd-iaai2022

References

Abdur Rehman, N.; Saif, U.; and Chunara, R. 2019. Deep landscape features for improving vector-borne disease prediction. In *CVPR Workshops*, 44–51.

Ayush, K.; Uzkent, B.; Burke, M.; Lobell, D.; and Ermon, S. 2020. Generating interpretable poverty maps using object detection in satellite images. In *IJCAI*, 4410–4416.

Brown, M. E.; Grace, K.; Shively, G.; Johnson, K. B.; and Carroll, M. 2014. Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. *Population and environment*, 36(1): 48–72.

Buchhorn, M.; Lesiv, M.; Tsendbazar, N.-E.; Herold, M.; Bertels, L.; and Smets, B. 2020. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6): 1044.

CIESIN. 2017. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11.

Elvidge, C. D.; Baugh, K.; Zhizhin, M.; Hsu, F. C.; and Ghosh, T. 2017. VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21): 5860–5879.

Gadiraju, K. K.; Ramachandra, B.; Chen, Z.; and Vatsavai, R. R. 2020. Multimodal Deep Learning Based Crop Classification Using Multispectral and Multitemporal Satellite Imagery. In *KDD*, 3234–3242.

Giglio, L.; and LANCE FIRMS. 2016. MODIS Aqua & Terra 1 km Thermal Anomalies and Fire Locations V006 NRT.

Golden, C. D.; Rice, B. L.; Randriamady, H. J.; Vonona, A. M.; Randrianasolo, J. F.; Tafangy, A. N.; Andrianantenaina, M. Y.; Arisco, N. J.; Emile, G. N.; Lainandrasana, F.; Mahonjolaza, R. F. F.; Raelson, H. P.; Rakotoarilalao, V. R.; Rakotomalala, A. A. N. A.; Rasamison, A. D.; Mahery, R.; Tantely, M. L.; Girod, R.; Annapragada, A.; Wesolowski, A.; Winter, A.; Hartl, D. L.; Hazen, J.; and Metcalf, C. J. E. 2020. Study Protocol: A Cross-Sectional Examination of Socio-Demographic and Ecological Determinants of Nutrition and Disease Across Madagascar. *Frontiers in Public Health*, 8: 500.

Hansen, M. C.; Potapov, P. V.; Moore, R.; Hancher, M.; Turubanova, S. A.; Tyukavina, A.; Thau, D.; Stehman, S.; Goetz, S. J.; Loveland, T. R.; et al. 2013. High-resolution global maps of 21st-century forest cover change. *science*, 342(6160): 850–853.

Huang, J.; Gretton, A.; Borgwardt, K.; Schölkopf, B.; and Smola, A. 2006. Correcting sample selection bias by unlabeled data. *NIPS*, 19: 601–608.

Humanitarian Data Exchange. 2020. Madagascar Healthsites. https://data.humdata.org/dataset/madagascar-healthsites. Accessed: 2021-12-20.

Ickowitz, A.; Powell, B.; Salim, M. A.; and Sunderland, T. C. 2014. Dietary quality and tree cover in Africa. *Global Environmental Change*, 24: 287–294.

International Food Policy Research Institute. 2020. Spatially-Disaggregated Crop Production Statistics Data in Africa South of the Sahara for 2017.

Author Manuscript

This article is protected by copyright. All rights reserved.

Keeley, B.; Little, C.; and Zuehlke, E. 2019. The State of the World's Children 2019: Children, Food and Nutrition–Growing Well in a Changing World. *UNICEF*.

Koppmair, S.; Kassie, M.; and Qaim, M. 2017. Farm production, market access and dietary diversity in Malawi. *Public health nutrition*, 20(2): 325–335.

Kroodsma, D. A.; Mayorga, J.; Hochberg, T.; Miller, N. A.; Boerder, K.; Ferretti, F.; Wilson, A.; Bergman, B.; White, T. D.; Block, B. A.; et al. 2018. Tracking the global footprint of fisheries. *Science*, 359(6378): 904–908.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *NeurIPS*, 30: 4765–4774.

McNally, A.; Arsenault, K.; Kumar, S.; Shukla, S.; Peterson, P.; Wang, S.; Funk, C.; Peters-Lidard, C. D.; and Verdin, J. P. 2017. A land data assimilation system for sub-Saharan Africa food and water security applications. *Scientific data*, 4(1): 1–19.

Micha, R.; Mannar, V.; Afshin, A.; Allemandi, L.; Baker, P.; Battersby, J.; Bhutta, Z.; Chen, K.; Corvalan, C.; Di Cesare, M.; et al. 2020. 2020 Global nutrition report: action on equity to end malnutrition. *Development Initiatives*.

Nakalembe, C. 2020. Urgent and critical need for sub-Saharan African countries to invest in Earth observation-based agricultural early warning and monitoring systems. *Environmental Research Letters*, 15(12): 121002.

NASA GSFC HSL. 2018. FLDAS Noah Land Surface Model L4 Global Monthly 0.1 x 0.1 degree (MERRA-2 and CHIRPS) V001.

NASA JPL. 2020. NASADEM Merged DEM Global 1 arc second V001 [Data set], NASA EOSDIS Land Processes DAAC. Accessed: 2020-12-30.

Park, H.-S.; and Jun, C.-H. 2009. A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*, 36(2): 3336–3341.

Pekel, J.-F.; Cottam, A.; Gorelick, N.; and Belward, A. S. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633): 418–422.

Poortinga, A.; Nguyen, Q.; Tenneson, K.; Troy, A.; Saah, D.; Bhandari, B.; Ellenburg, W. L.; Aekakkararungroj, A.; Ha, L.; Pham, H.; et al. 2019. Linking earth observations for assessing the food security situation in Vietnam: a landscape approach. *Frontiers in Environmental Science*, 7: 186.

Rasolofoson, R. A.; Hanauer, M. M.; Pappinen, A.; Fisher, B.; and Ricketts, T. H. 2018. Impacts of forests on children's diet in rural areas across 27 developing countries. *Science advances*, 4(8): eaat2853.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *KDD*, 1135–1144.

Robinson, T. P.; Wint, G. W.; Conchedda, G.; Van Boeckel, T. P.; Ercoli, V.; Palamara, E.; Cinardi, G.; D'Aietti, L.; Hay, S. I.; and Gilbert, M. 2014. Mapping the global distribution of livestock. *PloS one*, 9(5): e96084.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Shi, Z. R.; Wang, C.; and Fang, F. 2020. Artificial Intelligence for Social Good: A Survey. *arXiv preprint arXiv:2001.01818*.

Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; and Lucas, R. 2014. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sensing of environment*, 155: 13–31.

Steyn, N. P.; Nel, J. H.; Nantel, G.; Kennedy, G.; and Labadarios, D. 2006. Food variety and dietary diversity scores in children: are they good indicators of dietary adequacy? *Public health nutrition*, 9(5): 644–650.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI*, volume 30.

Sunderland, T.; O'Connor, A.; et al. 2020. Forests and food security: a review. *CAB Reviews*, 15(019): 1–10.

von Grebmer, K.; Saltzman, A.; Birol, E.; Wiesman, D.; Prasai, N.; Yin, S.; Yohannes, Y.; Menon, P.; Thompson, J.; Sonntag, A.; et al. 2014. 2014 Global Hunger Index: The challenge of hidden hunger. *IFPRI books*.

Xiong, J.; Thenkabail, P.; Tilton, J.; Gumma, M.; Teluguntla, P.; Congalton, R.; Yadav, K.; Dungan, J.; Oliphant, A.; Poehnelt, J.; Smith, C.; and Massey, R. 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Africa 30 m V001.

Author Bios

Dr. Elizabeth Bondi-Kelly is currently a Postdoctoral Fellow at MIT through the CSAIL METEOR Fellowship and an incoming Assistant Professor of Electrical Engineering and Computer Science at the University of Michigan. She has a PhD in Computer Science at Harvard University, where she was advised by Prof. Milind Tambe. Her research interests include multi-agent systems, remote sensing, computer vision, and machine learning, especially applied to conservation and public health. She has been recognized as an MIT EECS Rising Star in 2021, and has been awarded the Best Paper Runner Up at AAAI 2021, Best Application Demo Award at AAMAS 2019, Best Paper Award at SPIE DCS 2016, and an Honorable Mention for the NSF Graduate Research Fellowship Program in 2017. She has also founded Try AI, a nonprofit devoted to increasing diversity, equity, and inclusion in the field of AI.

Dr. Haipeng Chen is an assistant professor of data science at William & Mary. Previously, he was a CRCS postdoctoral fellow at Harvard University (with Milind Tambe), and a postdoc fellow in the Computer Science Department at Dartmouth College (with V.S. Subrahmanian). He obtained his Ph.D. from Interdisciplinary Graduate School (IGS), Nanyang Technological University (NTU), advised by Bo An and Yeng Chai Soh. He got the B.S. in Physics from University of Science and Technology of China (USTC). He has published in premier AI/data science conferences such as AAAI, IJCAI, NeurIPS, AAMAS, UAI, KDD, ICDM, and journals (e.g., IEEE/ACM Transactions, Transportation Research). He has published in premier AI/data science confer-

Author Manuscript

This article is protected by copyright. All rights reserved.

ences such as AAAI, IJCAI, NeurIPS, AAMAS, UAI, KDD, ICDM, and journals (e.g., IEEE/ACM Transactions, Transportation Research).

Dr. Christopher Golden is an Assistant Professor of Planetary Health and Nutrition at the Harvard T.H. Chan School of Public Health. As an ecologist and epidemiologist, his research investigates the human health impacts of global environmental change, with a focus on food systems. He received his BA from Harvard College and two graduate degrees from UC Berkeley: an MPH in Epidemiology with a focus in Nutrition, and a PhD in Environmental Science, Policy and Management. Golden has been conducting research in Madagascar for the past 23 years, and founded the non-profit Madagascar Health and Environmental Research (MAHERY). He has recently begun research in West Africa and the South Pacific. He is a core member of the CBD-WHO task force on biodiversity and health and the co-lead of the Nutrition chapter for the Blue Foods Assessment. His research has been published in Nature, Science, and the Proceedings of the National Academy of Sciences. His current research focuses on: 1) the role of climate-smart fisheries management to improve human nutrition; and 2) creating systems of climate-smart public health through climate and environmental monitoring and disease surveillance.

Nikhil Behari is a Research Support Associate at the MIT Media Lab in the Camera Culture group. He completed his A.B. at Harvard University. His research interests include computer vision and machine learning for sustainability and public health.

Dr. Milind Tambe is Gordon McKay Professor of Computer Science and Director of Center for Research in Computation and Society at Harvard University; concurrently, he is also Principal Scientist and Director "AI for Social Good" at Google Research. He is recipient of the IJCAI (International Joint Conference on Artificial Intelligence) John McCarthy Award, AAAI (Association for Advancement of Artificial Intelligence) Feigenbaum Prize, AAAI Robert S. Engelmore Memorial Lecture Award, AAMAS ACM (Association for Computing Machinery) Autonomous Agents Research Award, INFORMS (Institute for Operations Research and the Management Sciences) Wagner prize for excellence in Operations Research practice and Rist Prize from MORS (Military Operations Research Society). He is a fellow of AAAI and ACM. For his work on AI and public safety, he has received Columbus Fellowship Foundation Homeland security award and commendations and certificates of appreciation from the US Coast Guard, the Federal Air Marshals Service and airport police at the city of Los Angeles.

Author Manuscript

This article is protected by copyright. All rights reserved.