Data & Policy (2025), 7: e52
doi:10.1017/dap.2025.10013

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Assessing the utility of machine learning for predicting food sufficiency: a case study in Malawi

Andrew Tomes[1], Shahrzad Gholami[2], Didier Alia[1], Conor Hennessy[1], Dafeng Xu[1], Cecilia Bitz[3], Rahul Dodhia[2], Juan Lavista Ferres[2] and C. Leigh Anderson[1]

[1]Evans School of Public Policy and Governance, University of Washington, Seattle, WA, USA
[2]Microsoft AI for Good Lab
[3]Atmospheric Sciences Department, University of Washington, Seattle, WA, USA
Corresponding author: Andrew Tomes; Email: altomes@uw.edu

Received: 21 August 2024; **Revised:** 05 March 2025; **Accepted:** 30 May 2025

Keywords: crop prices; food insufficiency; machine learning; Malawi; remote sensing

Abstract

This study explores the potential of applying machine learning (ML) methods to identify and predict areas at risk of food insufficiency using a parsimonious set of publicly available data sources. We combine household survey data that captures monthly reported food insufficiency with remotely sensed measures of factors influencing crop production and maize price observations at the census enumeration area (EA) in Malawi. We consider three machine-learning models of different levels of complexity suitable for tabular data (TabNet, random forests, and LASSO) and classical logistic regression and examine their performance against the historical occurrence of food insufficiency. We find that the models achieve similar accuracy levels with differential performance in terms of precision and recall. The Shapley additive explanation decomposition applied to the models reveals that price information is the leading contributor to model fits. A possible explanation for the accuracy of simple predictors is the high spatiotemporal path dependency in our dataset, as the same areas of the country are repeatedly affected by food crises. Recurrent events suggest that immediate and longer-term responses to food crises, rather than predicting them, may be the bigger challenge, particularly in low-resource settings. Nonetheless, ML methods could be useful in filling important data gaps in food crises prediction, if followed by measures to strengthen food systems affected by climate change. Hence, we discuss the tradeoffs in training these models and their use by policymakers and practitioners.

Policy significance statement

Food insecurity continues to challenge Malawi's development efforts. Recent advancements in machine learning and the increased availability of high-frequency spatial and market data provide an opportunity to enhance short- and long-term policy responses through rapid predictions of food crises. While several studies have developed machine learning models for predicting food insecurity, these models remain complex and data intensive, creating barriers to uptake and implementation. We find that parsimonious models that leverage only public data on the production environment and prices achieve good accuracy in predicting food insufficiency in Malawi. We also find, however, that non-modeling predictive methods can have similar accuracy levels, although they present a tradeoff across error types. Overall, our findings suggest that machine learning models can be simplified

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

to be more accessible to policymakers, but that they may not outperform simpler heuristic approaches in predicting certain development outcomes, and their value rests with the existence of viable and implemented options for responding.

1. Introduction

Food access remains uneven across and within countries, and many low- and middle-income regions face endemic hunger. The most recent estimates suggest that in 2023, one in eleven people worldwide faced hunger and one in five people in Africa, where hunger is still on the rise (FAO et al., 2024). For households facing any period of moderate or severe food insufficiency, as measured by surveys on food availability, missed meals, and hunger over a 12-month period, this proportion has trended from 45.4% of the population in 2015 to almost 60% in 2023 (FAO et al., 2024). Hunger can be chronic or episodic and transitory if a household experiences and then recovers from a shock to production or income. However, hunger can also predictably recur if shocks are experienced on a seasonal basis. Seasonal hunger is transient but also cyclic, driven by intra-annual variability in food production and climate (Ayalew, 1997; Vaitla et al., 2009; Barrett, 2010; Gebrehiwot and Van der Veen, 2014; Anderson et al., 2017; Bonuedi et al., 2022).

Among rural populations, seasonal hunger food distribution inefficiencies arise from production variability, inadequate storage and market infrastructure, and shocks to international food supply and demand. Domestic policy also plays a role, as governments may intervene, for example, to alter prices, influence domestic supply via import or export controls, or influence production. Within a country, the political attention a region receives from the national government may be influenced by its agroecological endowment, with more productive areas receiving greater investments (Khandker and Mahmud, 2012). Policy approaches including production support (Madsden et al., 2021) and subsidies (Vaitla et al., 2009) have been impactful, but food insufficiency and calls for improved guidance remain (Vaitla et al., 2009).

There has been substantial progress developing tools, including using machine learning (ML) methods, to anticipate food insufficiency. This work is supported by improvements in the availability of diverse data, including surveys (Gholami et al., 2022), news reports (Balashankar et al., 2023), and mixtures of survey-sourced and satellite-imaged variables (Lentz et al., 2019). There still exist, however, challenges to making these models usable by governments or civil society organizations including the complexity of the methods and costly data requirements. Furthermore, while anticipating crises can be useful to avert severe mortality and morbidity, cumulative mortality can be higher at lower severity levels of food insufficiency because they can persist for longer (Maxwell et al., 2020). The transition into what becomes an officially declared famine is not necessarily always the abrupt change that is typically the focus of transition models (see Westerveld et al., 2021 and Krishnamurthy R et al., 2022), suggesting that ongoing monitoring is needed to supplement anticipatory modeling.

In this paper, we ask whether relatively easily collected information on crop production conditions and prices can be reliably used to develop models that identify and predict food insufficiency at the scale of a census enumeration area (EA). We focus on Malawi, a country subject to recurring food crises and where we have rich nationally representative georeferenced household survey data on reported monthly experience of food insufficiency. Between 2010 and 2020, Malawi experienced five major food crises caused by climate and energy price shocks (Supplementary Figure A2). The crises are associated with substantial increases in the proportion of households reporting food insufficiency during the crisis years and afterward. Our analysis aims to assess whether public and high-frequency remotely sensed and price data can be used with ML to detect food insufficiency. We examine the tradeoffs in modeling choices when training these models to facilitate a greater understanding of their performance. Finally, even with a parsimonious modeling approach, we also ask whether ML methods sufficiently improve accuracy compared to simpler non-modeling approaches.

Our study contributes to an emerging and rapidly growing literature on applying ML to development challenges by looking more deeply into model and data tradeoffs. In their 2019 study, Lentz and others use a combination of readily available data on rainfall quantity and seasonal variation, soil quality, market

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

prices, and access to market (termed "Class 1" data) combined with more difficult-to-collect data on household assets (such as roofing material, termed "Class 2") and household characteristics ("Class 3") gathered via survey. They find that LASSO regression anticipates, with roughly 80% success, future binned EA-level averages of the household dietary diversity score (HDDS), one of three continuous measures of food insufficiency. Model fits were less successful for the reduced coping strategies index (rCSI) and food consumption score (FCS), a nutrition-weighted index of dietary diversity. Similar work by Zhou et al. (2022) also find roughly 60–70% accuracy in modeling on FCS and rCSI in Malawi waves 1–3, with time-lagged prices as a primary driver.

Finally, a study was conducted by Gholami et al. (2022) using high-frequency survey data collected between 2017 and 2019 in southern Malawi to estimate categorized rCSI scores from household characteristics, including location (longitude/latitude), beliefs about future welfare, and experienced shocks. The models were trained using random forests (RFs) and had out-of-sample and forward predictive (one month in advance) accuracies of 70–80%. Model decomposition suggested that household location and subjectively assessed current and anticipated welfare made the largest contributions.

Our analysis extends Lentz et al. (2019) and Zhou et al. (2022)'s data by adding the third and fourth waves of Malawi's Integrated Household Survey Program (IHS) of the Living Standards Measurement Study—Integrated Survey on Agriculture (LSMS-ISA). The inclusion of these additional data improves the ability to evaluate out-of-sample accuracy, given that the second wave of the Malawi LSMS-ISA was a panel subset of the first wave, while waves 3 and 4 contain nationally representative cross sections. Our focus is on building a more parsimonious model and understanding the gains of modeling complexity. We therefore begin by comparing sets of predictors and methods. For predictors, we compare price, inflation, remotely sensed variables, and their combination to a benchmark consisting of month and region. For methods, we compare three ML models of different complexities (ridge regression, RFs, and TabNet neural networks) against a benchmark using classical logistic regression. We also applied Shapley additive explanations (SHAP) techniques explored in Gholami et al. (2022) and Martini et al. (2022) to assess the relative contributions of remotely sensed and price data to predictive accuracy.

We find that the various models produce comparable levels of acceptable accuracy, though they differ in precision and recall. Additionally, we note that performance is influenced by modeling decisions, such as how cutoff points are defined to categorize an EA as food insufficient and how the training set is sampled across different waves. The application of SHAP decomposition shows that nominal maize prices, their trends, and the broader consumer price index (CPI) inflation rate—which reflects the combination of supply and demand factors across rural producers and urban consumers—are the most significant factors contributing to predictive accuracy. However, we also find that the improvements in accuracy provided by ML models over simpler, less computationally demanding prediction methods are minimal. Given the increasing popularity of ML techniques and their potential to address important data, we discuss tradeoffs involved in training these models and their use in food security prediction by policymakers and practitioners. We conclude that despite promising results on the accuracy of simpler ML models, particularly in regions where the spatial and temporal variability of climate shocks is limited, non-modeled approaches may be sufficient. Successfully responding both immediately to food crises and addressing the underlying contributors over the longer term, rather than successful prediction, may be the bigger challenge, particularly in low-resource settings.

2. Data and methods

2.1. LSMS-ISA IHS survey data

Data on household experience of food insufficiency come from Malawi's IHS supported by LSMS-ISA. This national panel survey is conducted every two to three years. The LSMS-ISA in Malawi extends the IHS beginning in 2010, when the full instrument was first implemented for a panel of 4,000 households within the larger IHS sample. This study uses all mainland Malawian households of the wave 1 IHS sample (12,271 households, 2010–11), the panel only IHPS/LSMS-ISA Wave 2 survey whose locations

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

remained close to their original wave 1 locations (3,385 households, 2013–14), and the combined IHS/LSMS-ISA Wave 3 and Wave 4 cross-sectional surveys (12,191 and 11,250 in 2015–16 and 2018–19, respectively).

Our outcome variable is based on responses to the survey question asking respondents to list which months out of the previous twelve the household "ran out of food." Previous studies have used alternative food security measures such as the Food Consumption Score (FCS), the HDDS, and the Reduced Coping Strategy Index (rCSI) (see Supplementary Material, Appendix C for full discussion). These indicators are derived using questions that rely on a short recall period, typically capturing household experiences over the past seven days—hence, they are limited in their ability to account for seasonality in households' experience of food insufficiency. In contrast, the "ran out of food" variable captures households' experience over 12 months. Although this longer recall period may increase the risk of recall error, it allows a more comprehensive view of households' experience of food insufficiency and is more suitable for capturing seasonal fluctuations. Responses were used to generate a binary variable based on thresholds for the average proportion of households with insufficient food in each EA by month and survey wave. We drop EA–month pairs with fewer than eight observations. We also drop observations with inconsistencies in data entry between the cross section and the panel sample in waves 3 and 4. In wave 2, households that were reported as having moved more than 10 km away from their previous IHS location were also excluded, because their original EA ID was not updated. With these exclusions, there are approximately 700 EAs per wave except for wave 2 (204 EAs), with a total of 2,440 across all four waves. The number of households per EA ranged from 8 to 26, with the majority (2,217) consisting of a sample of 16. The final sample consists of 29,320 EA–month observational pairs. Food insufficiency rates are shown in Figure 1. A detailed explanation of response variable preparation is provided in Supplementary Material, Appendix A.

EAs were classified by their proportion of food-insufficient households, under the assumption that higher prevalence implied greater severity with fewer opportunities to share food. All EAs experiencing some food insufficiency were divided into quartiles. In Table 1, category 1 (c_1) is the 3rd quartile of all insufficiency observations and corresponds to two to nine or an average of just under one-third of households experiencing food insufficiency. Category 2 (c_2) is based on the fourth quartile and has a

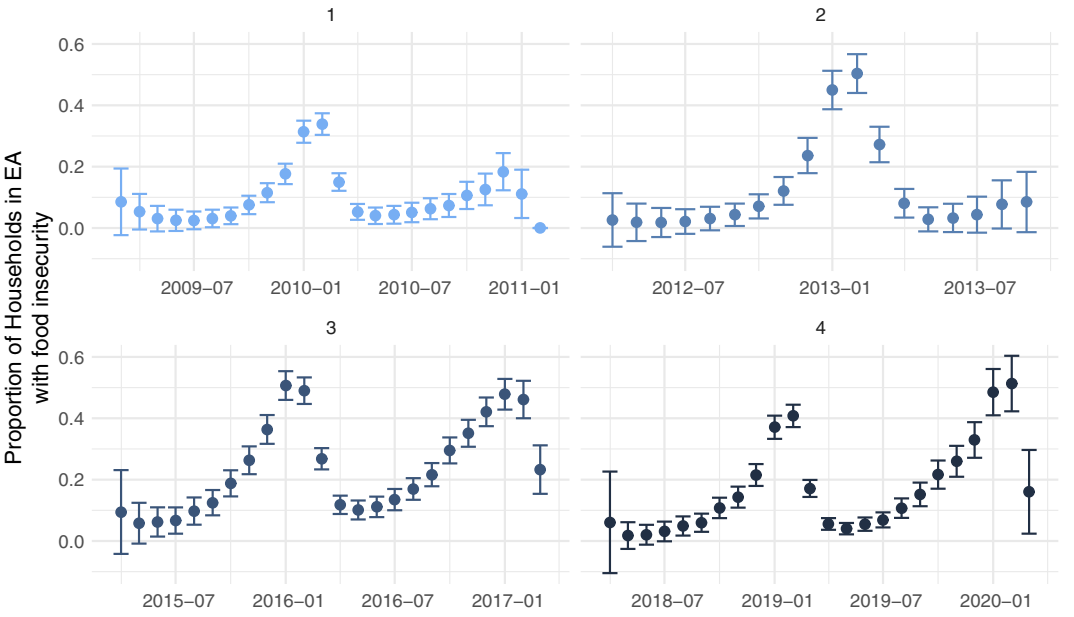

Figure 1. *Average proportion of households with insufficient food in each EA by month and survey wave (subplot title). Error bars represent the 95% CI.*

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

Table 1. *Classification categories and their compositions*

Category	*N*	Percent of dataset	Minimum number of food-insufficient households	Average proportion of food-insufficient households	Maximum number of food-insufficient households
0	20,245	69%	0	5.6%	4
1	5,687	19.4%	2	32.7%	9
2	3,388	11.6%	4	62.9%	20

Note: Category 1 (c_1) covers the upper third quartile of food insufficiency and includes EAs with at least two households reporting a food shortage.

minimum of nearly 40% of households experiencing food insufficiency. These categories are derived from observations from the data itself rather than externally imposed. This approach reduces the lumping of disparate EAs into the same category (see Supplementary Material, Appendix C), but greater imbalance across categories can reduce classification accuracy. In this case, accuracy at the lower threshold for classification might be higher, but the predictions may be less useful as they cover a wider range of insufficiency levels.

2.2. Market price information

Price data are taken from the UN's World Food Programme market observers. The observations are collected monthly by local observers at marketplaces. Prices are not available for some month–market combinations, with gaps of 1–3, and rarely up to 7 months. Lentz et al. (2019) and Zhou (2020) consider these missing observations an indication of "market thinness." However, missing observations sometimes result from observers being unavailable (and households do not often cite lack of food availability in the market; see Supplementary Material, Appendix C), and so we instead linearly interpolate using prices on either side of the gap (as in Andrée, 2021). The point observations are then spatially interpolated to generate distance-weighted price gradients across markets.

2.3. Remotely sensed data

The model incorporates modeled datasets from the TerraClimate database and NASA's moderate resolution imaging spectroradiometer (MODIS). TerraClimate is a derivative of the WorldClim dataset, which is constructed using spatial interpolation on ground-based station measurements (9,000 to 60,000 depending on year; see Fick and Hijmans, 2017) using MODIS observations for additional information (Abatzoglou et al., 2018). The TerraClimate variables include monthly Palmer Drought Severity Index (PDSI), a relative scale that ranges from −10 (extremely dry) to +10 (extremely wet); monthly precipitation; monthly mean soil moisture; monthly mean vapor pressure; and monthly mean vapor pressure deficit. The MODIS datasets include land cover classification, surface temperature, and the normalized difference vegetation index (NDVI). The annual land cover classification is from the MCD12Q1 modeled dataset (Friedl and Sulla-Menashe, 2019). Daily daytime minimum, maximum, and mean surface temperature from the MOD11A1 dataset (Wan et al., 2015) are averaged to generate monthly mean, monthly mean minimum, and monthly mean maximum temperature for analysis. The 16-day NDVI from MOD13Q1 (Didan, 2015) is converted to monthly by averaging (see Figure 2).

2.4. Data processing

All MODIS and TerraClimate spatial datasets are processed using rioxarray and xarray packages in Python. The retrieved datasets are first resampled to a standard 250-m grid. For each variable, a 20-by-20-pixel square (25 km^2) containing the EA is extracted and averaged (or, for land cover, the total number of tiles belonging to each class is counted) to produce a set of observations. The remotely sensed imagery covers the most recent growing season (running from April to March) prior to the reference

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

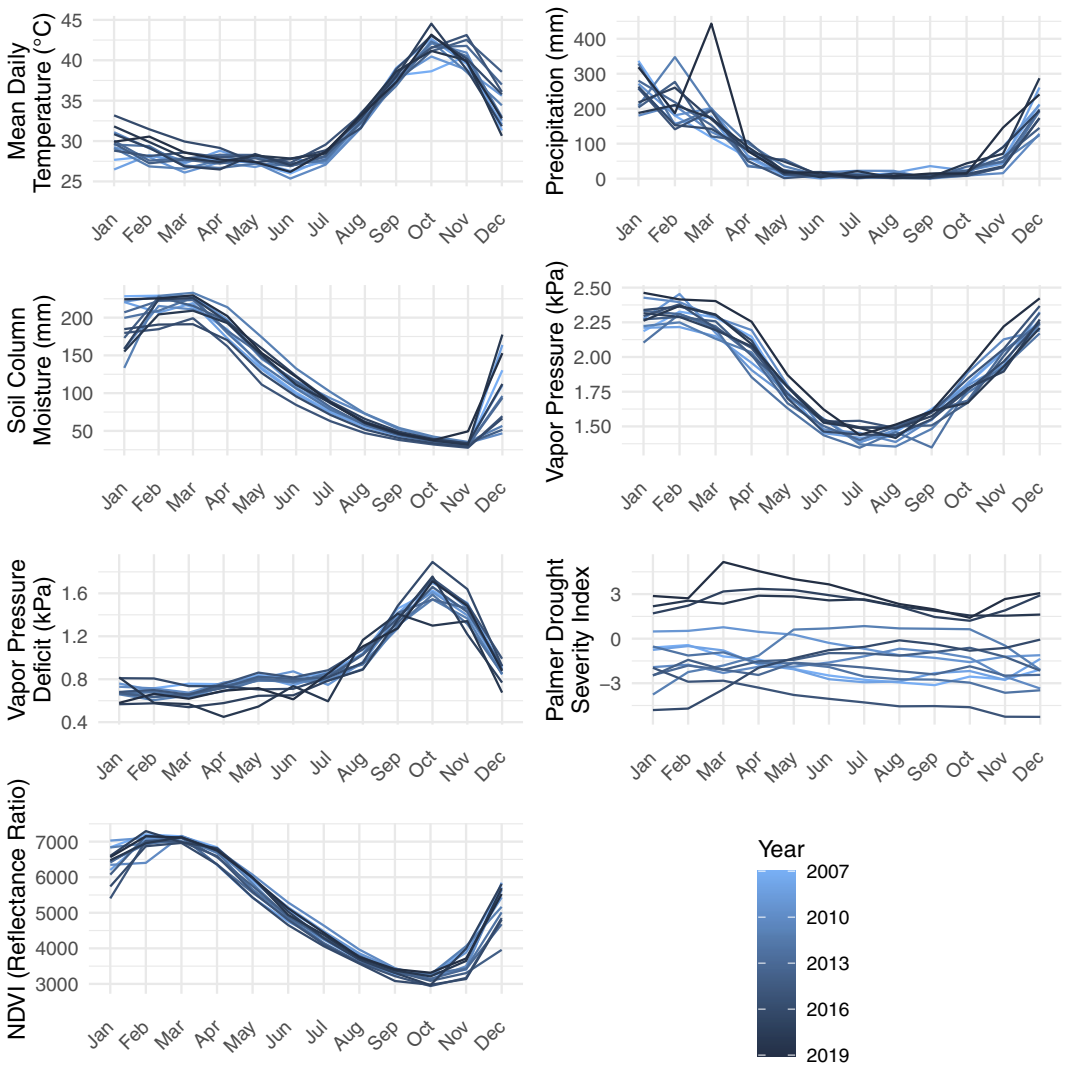

Figure 2. *Selected spatial data means across EAs for the time range included in this study. From upper left to lower right, mean surface temperature, precipitation, soil moisture, vapor pressure, vapor pressure deficit, Palmer Drought Severity Index (PDSI), and vegetation index (NDVI).*

month (see Supplementary Material, Appendix A). For maize price data, we calculate the nominal price and relative year-over-year change (monthly inflation) for each of the twelve months prior to the reference month. We also include a measure of recent price movements, the relative change between the average price of the current quarter and the average price of the previous quarter.

2.5. Model fitting and evaluation

We try two approaches to the classification of food insufficiency. The first assumes the data has been collected with partial spatial coverage. The model goal is then to interpolate—that is, to attempt to assess the food insufficiency in areas or time periods not sampled. We attempt this process for waves 1, 3, and 4, omitting wave 2 due to the small sample sizes (less than a third of each of the other waves). The second approach is a more traditional attempt at forecasting, using data from prior survey rounds to estimate the

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

food insufficiency occurrence in the future. A key decision in ML modeling using longitudinal data involves the choice of data partitioning in training and test sets. Bergmeir and Benitez (2012) describe four data partitioning techniques: fixed-origin, rolling-origin-recalibration, rolling-origin-update, and rolling-window evaluation. With fixed-origin partitioning, forecasts are made originating at the point subsequent to the last one in the training data. Rolling-origin-recalibration sequentially moves values from the test set to the training set, and rolling-origin-update changes the forecast horizon without updating the training data. The advantage of the latter is that the model only needs to be trained a single time, whereas the former must be retrained as new data are added and requires ongoing maintenance. Finally, rolling-window evaluation resembles rolling-origin-update, except that old training data are dropped from the model as new data are added. The advantage of rolling-window evaluation is that it can keep older, potentially less relevant data from biasing the model, while at the expense of potentially useful information being dropped. Lentz et al. (2019), Martini et al. (2022), and Zhou et al. (2022) use fixed-origin partitioning in their food security predictions. We select rolling-origin-recalibration, first training the model on survey wave 1, predicting on survey wave 2, training on waves 1 and 2 and predicting on wave 3, and training on waves 1 through 3 and predicting on wave 4, as this approach allows to test for differences in predictive accuracy as additional training data are introduced, simulating how this task would be approached in practice.

Testing on multiple subsamples of the existing data can be used to evaluate sampling error and test for overfitting. A common cross validation strategy is *k* fold cross validation, which splits the data into *k* subsets, training on each subset and testing on the remaining *k*-1 subsets. For experiment 1, we perform five-fold cross validation at a train:test ratio of 20:80, selecting on the EA. For model training in Experiment 2, we randomly sample 100 EA observations per month, which range from roughly 12.5% to 100% of the available monthly observations, with the lean season being more intensively sampled than the non-lean season (Figure 3). The testing is performed on the full set of observations from the subsequent wave. In initial trials, we included hyperparameter tuning but found that model classification

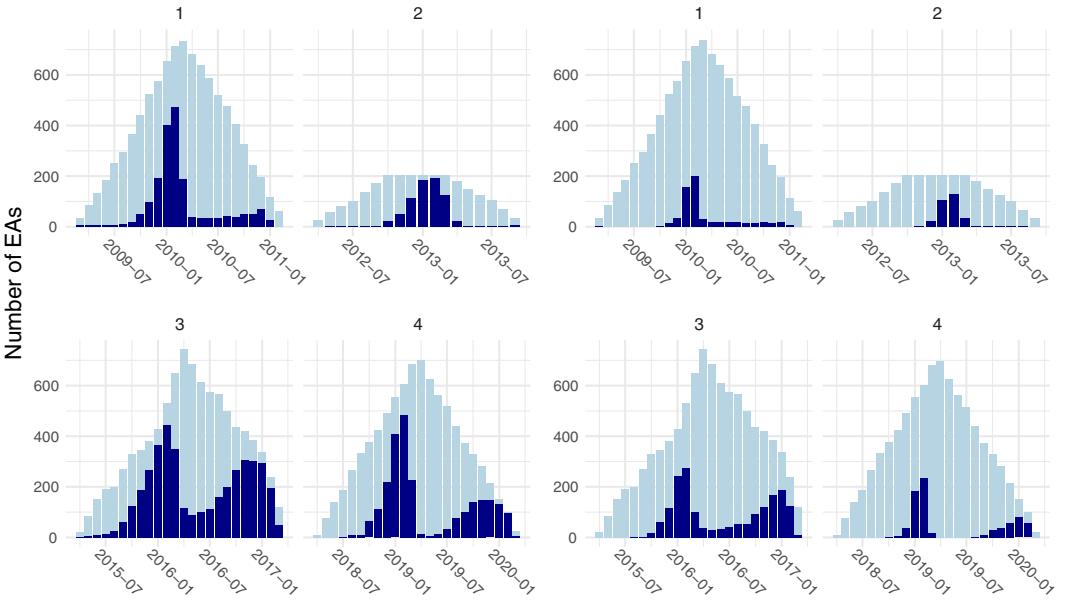

Figure 3. *Comparison of the two food insufficiency classifications and sampling intensity across the four survey waves. The c1 classification is shown on the left and c2 on the left. The dark bars indicate the number of food insufficient EAs while the light bars indicate the number of food sufficient EAs for each month.*

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

accuracy was not particularly sensitive to hyperparameters, and so results are reported based on the default settings of each model.

We consider three models suitable for tabular datasets—LASSO, RF, and TabNet, and logistic regression, which remains a classical approach to modeling binary responses and is commonly chosen as a benchmark for comparing the machine-learning models (Grinsztajn et al., 2022). LASSO regression is a linear regression method that drops variables with the least impact on model accuracy to maximize parsimony and is a relatively simple ML strategy. RF constructs a set of decision trees using a random subset of features and the training data and then uses then aggregates their predictions either using the majority vote or average to make the final prediction. TabNet, in contrast to regression trees, is a deep learning method that is optimized for tabular data (see Arik and Pfister, 2021). The models are computed in Python using the scikit-learn package v1.3.0.

The models are first trained on a minimal dataset consisting solely of the annual quarter of observation offset by one month to better align with the growing season, a regional dummy (1–3), and a rural dummy (0–1). To this specification, we add different variable sets (see Table 2); the "LSMS" set, consisting of distances to roads, markets, borders, local government offices, and the nearest population center, elevation, and proportion of the EA that is used for agriculture; the "remotely sensed" set consisting of the precipitation, drought severity, temperature, NDVI, and land cover variables; the "price" set consisting

Table 2. *Description of the variables used for each specification*

Variable set	Contains	Total number of predictors
Minimal	Rural dummy, quarter dummy (3), and region dummy (2)	6
Remotely sensed	Annual land cover classes (crop, urban, forest, grassland), monthly total precipitation and mean NDVI, PDSI, vapor pressure, vapor pressure deficit, and surface temperature (average of daily minimum, maximum, and average)	112
LSMS	Distances from the nearest road, population center, Agricultural Development and Marketing Corporation (ADMARC) center, auction, government offices, and border post; approximate coverage of EA footprint by agricultural area, and elevation	8
Price	Average monthly maize price for the previous 12 months	12
Inflation	Average monthly maize price inflation and CPI inflation	24

Note: All model specifications other than the minimal set include the minimal set for fixed effects in addition to the variables listed.

Table 3. *Summary statistics for selected variables; NDVI and PDSI are both indices and are unitless*

Pooled sample	Mean	Median	St. Dev.	Min	Max
Distance to road (km)	8.44	1	0.39	0	1
Elevation (m)	871	899	343	37	1754
Maize price inflation	0.26	0.29	0.515	−0.66	3.03
Maize price (Malawian kwacha [MWK])	133	133	87	17	361
Midseason NDVI	6304	6408	1027	886	8500
Midseason precipitation (mm)	251	252	82	82	630
Midseason PDSI	−0.49	−1.09	3.12	−7.81	6.69
Rural (0/1)	0.81	1	0.39	0	1

Note: For variables labeled "midseason," January was chosen as the approximate middle of the agricultural season.

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

of monthly price observations for the twelve months prior to the observation period; and the "inflation" set consisting of monthly maize price inflation and national CPI inflation. We compare combinations of these variables (prices + inflation, prices + LSMS, prices + remotely sensed, inflation + remotely sensed, and inflation + LSMS) to a regression tree-based selection algorithm applied to all variables (the "selected" set). Summary statistics for selected variables are presented in Table 3. At the c_2 classification, the comparatively small number of positive cases made it difficult for model fits to converge, so we applied synthetic minority oversampling technique (SMOTE) using the package imblearn. Synthetic oversampling creates additional positive cases based on observed positive cases, with some additional random noise added to the predictor variables.

2.6. Validation

The goal of this exercise is to use a computational model to classify EAs as either food sufficient (negative for food insufficiency) or food insufficient (positive for food insufficiency) according to the thresholds described in Table 1. There are several metrics for comparing the performance of classifiers. Accuracy, the ratio of correct positive and negative assignments to all assignments, is the most intuitive, but it can be misleadingly high in situations where the target classification is rare. In a scenario where the population consists of 90 negative observations and 10 positive observations, a classifier that assigns all observations to the most common class would have an accuracy of 90%. This finding would imply a highly functional model, but it would provide no utility for identifying the subpopulation of interest. Instead, precision and recall can be used to determine the model's skill in making true positive assignments. Precision is a measurement of the ratio of true positive assignments to all positive assignments (i.e., the inverse of the false negative rate). Recall is the ratio of true positive assignments to all positive cases (the true positive rate). Precision and recall are inversely related to one another; increasing precision typically comes at the expense of reducing recall and vice versa, and model training weights can be used to establish a relative preference of one over the other. The tradeoff between precision and recall can be measured in terms of F1, an average of precision and recall, and the area under the precision-recall curve (AUPRC), which is a function of the precision and recall at each possible threshold (on a scale of 0 to 1) used to determine which of the continuous set of fitted values produced by the model is a positive assignment and which is a negative assignment (see Table 4). Higher values indicate a better balance between precision and recall.

In addition to these metrics, we add some general principles that can be used to evaluate model performance and validity:

Models should outperform a "most frequent case classifier." Performance on preferred metrics should be higher than simply assigning all observations to the most abundant class in the dataset. For binary comparisons, this can be a trivial exercise, as precision and recall will always be higher when negative cases are most abundant, but, for multiple classes where more than one may be of interest to the modeler, this criterion can be useful for measuring performance.

Table 4. *Evaluation Metrics for the performance of a classifier*

Metric	Equation	Interpretation
Precision	$\frac{TP}{TP+FP}$	Number of true positive predictions (i.e., food-insecure households predicted as food insecure) out of all positive predictions
Recall	$\frac{TP}{TP+FN}$	Number of true positive predictions identified out of all positive cases
F1	$2 \times \frac{precision \times recall}{precision + recall}$	Harmonic mean of precision and recall
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	How well the algorithm has classified positive and negative classes over the total cases

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

More complex approaches should outperform simpler ones. Although data storage and computation time have become comparatively inexpensive, constructing and using ML models can still require substantial person-hours and at least the initial involvement of someone with expertise. Compiling and maintaining large datasets can require ongoing support. Thus, if a simpler method can perform similarly to a more complex one, it may be necessary to conclude that either the dataset is inadequate for the chosen task or that a more complex approach is unnecessary, depending on overall performance. Grinsztajn et al. (2022) quantify this principle in terms of "easiness" and conclude that a dataset is too easy when an ML approach fails to outperform a simpler one (such as OLS regression) by at least five percent on the chosen metrics. Here, we compare the ML approaches to logistic regression.

The number of predictor variables should be as small as possible: we define a parsimonious set of predictors of variables representing the EAs district, the time of year, and a rural-urban stratum. These metrics provide very coarse indicators of time of year and location, and it should be possible to outperform them with more geographically detailed information.

For predictive models, performance should be better than someone with simple knowledge of the past: We assume that practitioners and policymakers will have access to past information on where food insufficiency existed because it is impossible to create a model without training data. Consequently, if trends in food insufficiency are highly spatiotemporally consistent, it may be sufficient to know how well a community near a target community was doing in a previous survey. We refer to this approach as the "naive" approach and compare the accuracy, precision, and recall of predicting a given EA's food insufficiency based on the observed food insufficiency of its nearest neighbor in the previous survey round to the accuracy, precision, and recall of the modeled forward predictions.

3. Results

3.1. Experiment 1: Within-wave models

3.1.1. Results for the c_1 *category*

Across all experiments, the selected variable set tended to be close to the total variable set, indicating poor performance in finding the most informative variables. Results from that trial are excluded. In the within-wave models, excluding models that did not converge (precision or recall less than 0.1), accuracy scores ranged from 74% (TabNet fitting on wave 3 data using remotely sensed variables) to 87% (RFs fits on wave 1 data using either inflation and remotely sensed variables or all variables). Compared to the logistic regression, LASSO fits were identical and RFs and TabNet were within three to four percentage points in accuracy, but some results had substantially higher precision; e.g., the RFs fit was 13 percentage points higher than logistic regression on the inflation and remotely sensed set and all variables set in wave 1, producing the greatest observed AUPRC of 0.7. This difference was much lower in wave 3 (four points) and moderately lower in wave 4 (eight points). While recall was similar within variable sets, across sets, inflation had substantially higher recall in waves 3 and 4 (8- and 12-point increases over the minimal set, respectively), while it did not have a substantial effect alone in wave 1, while it produced a 7-to-8-point increase in combination with the remotely sensed variables. Figure 4 compares results for all variable sets and models against the minimal logistic regression, and Tables 5–7 present the most accurate fits for waves 1, 3, and 4 compared to the minimal logistic regression.

3.1.2. Results for the c_2 *category*

It was not possible to evaluate performance against the minimal specification for wave 1 fits because it is not possible to oversample on exclusively categorical predictors and the minimal fits failed to converge. Across the remaining specifications, performance was slightly higher for TabNet and RFs than for logistic regression. High accuracy scores were accompanied by low precision and recall, indicating a bias toward false negatives. Substantial increases in both precision and recall occurred in waves 3 and 4, with the price specification of TabNet achieving substantial improvements in recall (32.6 points compared to the

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

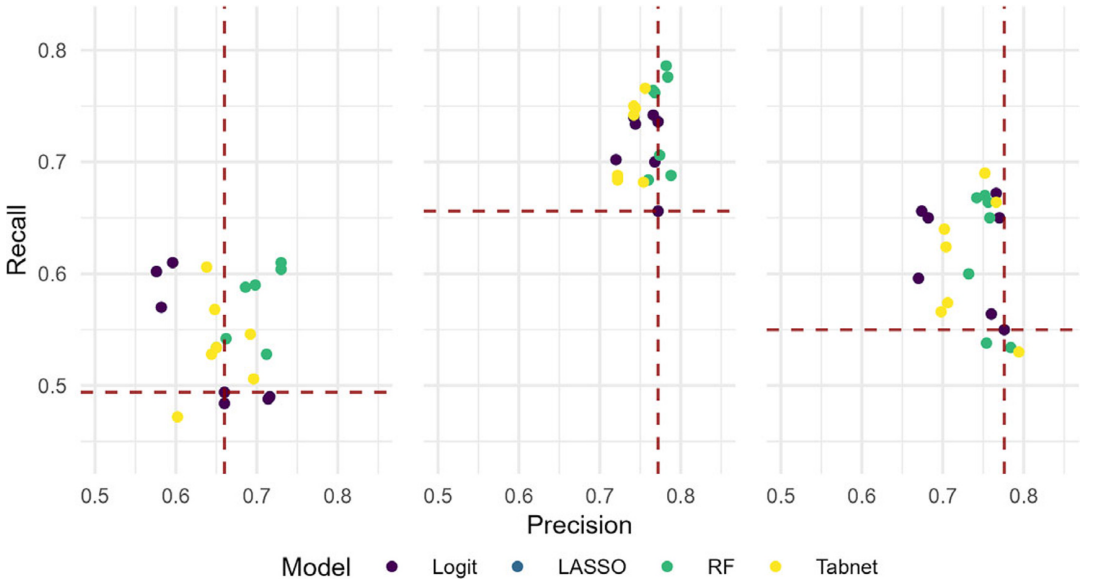

Figure 4. *From left to right, comparisons for waves 1, 3, and 4 in terms of precision and recall across model specifications for all models at the* c_1 *threshold using the within-wave training and testing datasets. The red dotted lines show the precision and recall fit with logistic regression and the minimal variable set. Points that fall in the upper right quadrant are more accurate, while those in the upper left outperform on recall but underperform on precision, and those in the lower right outperform on precision but underperform on recall. Abbreviations: logit: logistic regression, RF: random forests.*

Table 5. *Comparison of the simple logit fit to higher-performing models in wave 1 at the* c_1 *threshold*

Variable set	Model	Precision	Recall	Accuracy	AUPRC
Minimal	Logit	0.66	0.494	0.84	0.63
Inflation + remotely sensed	Random forest	0.73	0.604	0.874	0.774
All variables	Random forest	0.73	0.61	0.866	0.686
Price	Random forest	0.686	0.588	0.86	0.682
Remotely sensed	Random forest	0.712	0.528	0.86	0.67

Table 6. *Comparison of the simple logit fit to higher-performing models in wave 3 at the* c_1 *threshold*

Variable set	Model	Precision	Recall	Accuracy	AUPRC
Minimal	Logit	0.772	0.656	0.758	0.792
All variables	Random forest	0.782	0.786	0.808	0.832
Inflation + remotely sensed	Random forest	0.784	0.776	0.804	0.83
Inflation	Random forest	0.766	0.764	0.79	0.818
Price	Random forest	0.768	0.762	0.788	0.816

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

Table 7. *Comparison of the simple logit fit to higher-performing models in wave 4 at the* c_1 *threshold*

Variable set	Model	Precision	Recall	Accuracy	AUPRC
Simple	Logit	0.776	0.55	0.82	0.728
Inflation	LASSO	0.766	0.672	0.844	0.764
Inflation	Logit	0.766	0.672	0.844	0.764
Price	LASSO	0.77	0.65	0.842	0.76
Price	Logit	0.77	0.65	0.842	0.76

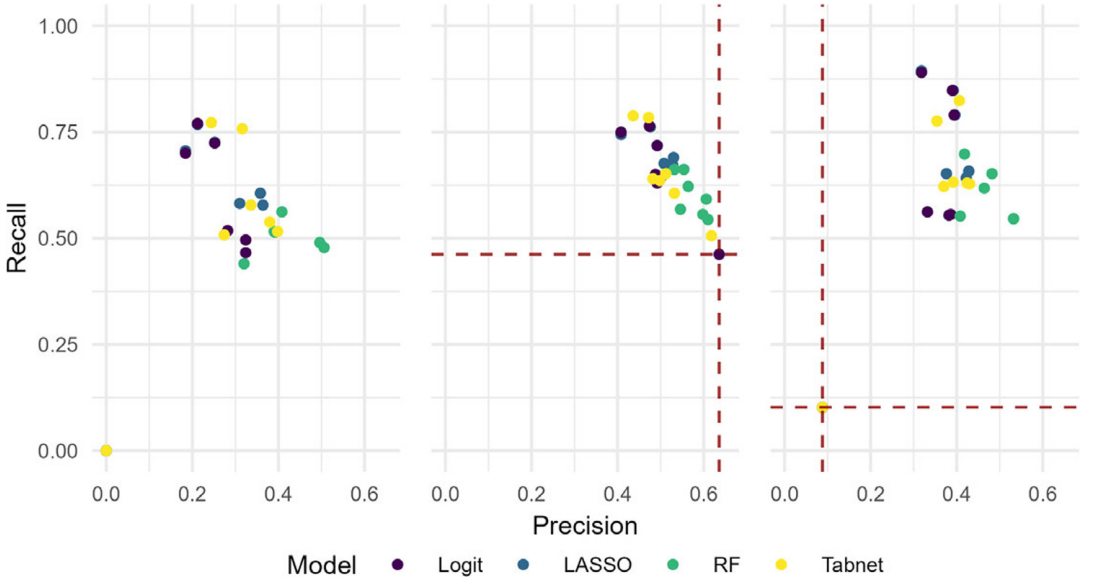

Figure 5. *Comparison of precision and recall across model specifications for all models at the* c_2 *classification using the within-wave training and testing datasets. Waves 1, 3, and 4 are presented from left to right.*

Table 8. *Comparison of the best-performing models to the minimal logit model in terms of classification accuracy, precision, and recall using the* c_2 *category in wave 1*

Variable set	Model	Precision	Recall	Accuracy	AUPRC
Minimal	Logit	0	0	0.938	0.5
Inflation + remotely sensed	Random forest	0.496	0.49	0.936	0.51
Price	Random forest	0.408	0.562	0.92	0.498
All variables	TabNet	0.398	0.516	0.918	0.474
Inflation	Random forest	0.392	0.514	0.916	0.47

Note: The zeros in precision and recall for the minimal logit model indicate that the model is functioning as a most common case classifier and did not produce any positive assignments.

minimal/logistic regression benchmark in wave 3 and 74.8 points in wave 4) with accompanying losses in precision (20 points below the minimal/logistic regression benchmark in wave 3 and 31 points above benchmark in wave 4, but below comparable RF scores by 19–20 points). Figure 5 compares the results across all variable sets and model fits to the benchmark, and Tables 8–10 present details of the best performing models compared to the benchmark for waves 1, 3, and 4.

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

Table 9. *Comparison of the best-performing models in wave 3 at the* c_2 *threshold*

Variable Set	Model	Precision	Recall	Accuracy	AUPRC
Minimal	Logit	0.636	0.462	0.842	0.602
All variables	Random forest	0.606	0.592	0.844	0.64
Minimal	Random forest	0.61	0.544	0.842	0.62
Inflation + remotely sensed	Random forest	0.598	0.556	0.84	0.622
Price	TabNet	0.436	0.788	0.784	0.65

Table 10. *Comparison of the best-performing models in wave 4 at the* c_2 *threshold*

Variable set	Model	Precision	Recall	Accuracy	AUPRC
Minimal	Logit	0.088	0.102	0.908	0.538
Inflation + remotely sensed	Random forest	0.532	0.546	0.916	0.56
Price	Random forest	0.482	0.652	0.906	0.584
Inflation	Random forest	0.564	0.618	0.902	0.56
Price	LASSO	0.392	0.848	0.864	0.626

3.1.3. SHAP decompositions for the random forest models

To understand the contributions of each variable to the predictive accuracy, we calculate SHAP values on the RF models. SHAP values show the observation-level contributions of each variable to the predicted value. A greater magnitude SHAP value indicates greater influence in determining the prediction, while signs indicate the direction of the influence—either toward the positive case (the EA is food insufficient) or away if negative. In the figures below, the SHAP value is on the X axis and variable values are illustrated through a color scale where red indicates highly positive values and blue indicates values that are highly negative or close to zero depending on the variable's scale. The variables, filtered to the most important 20 (or all if the model specification had fewer than 20 variables), are on the Y axis in descending order of average contribution to the predictions across the entire dataset. The points (representing individual EA/month observations) are jittered in the Y axis direction to get a sense of the distribution of the Shapley values per variable. A greater width in the band indicates a greater number of observations.

In Figure 6, we present SHAP values for the experiment 1 c_1 RF fits on all variables. Across all waves, inflation, nominal maize prices, and maize inflation tended to rank highly, along with the temporal and spatial dummies. In waves 1 and 3, precipitation and vapor pressure deficit (vpd) during the growing season are also influential, with greater vpd and lower precipitation values associated with a greater likelihood of food insufficiency, likely corresponding with the greater aridity occurring in the southern region. Results for the c_2 classifier were similar (Supplementary Material, Appendix B)

3.2. Experiment 2: Forward predictions

3.2.1. Benchmarking with the non-model approach

To establish the benchmarks, a non-model method where the predicted value for each EA in each survey wave was taken from its nearest neighbor (determined based on Euclidean distance between the centroid coordinates provided in the LSMS-ISA) in the previous wave. For each wave, the neighbor value from the previous wave was compared to the observation value for the current wave to generate estimates of accuracy, precision, and recall. Because the wave 2 survey was a panel subset of the wave 1 survey, we remove the panel EAs from wave 1 to avoid the equivalent of overfitting. Comparing the classification accuracy of the non-model estimator to the accuracy of simply using the state of the EA in the previous survey provides an estimate of the effect of including the panel in the modeling approaches. For the c_1

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

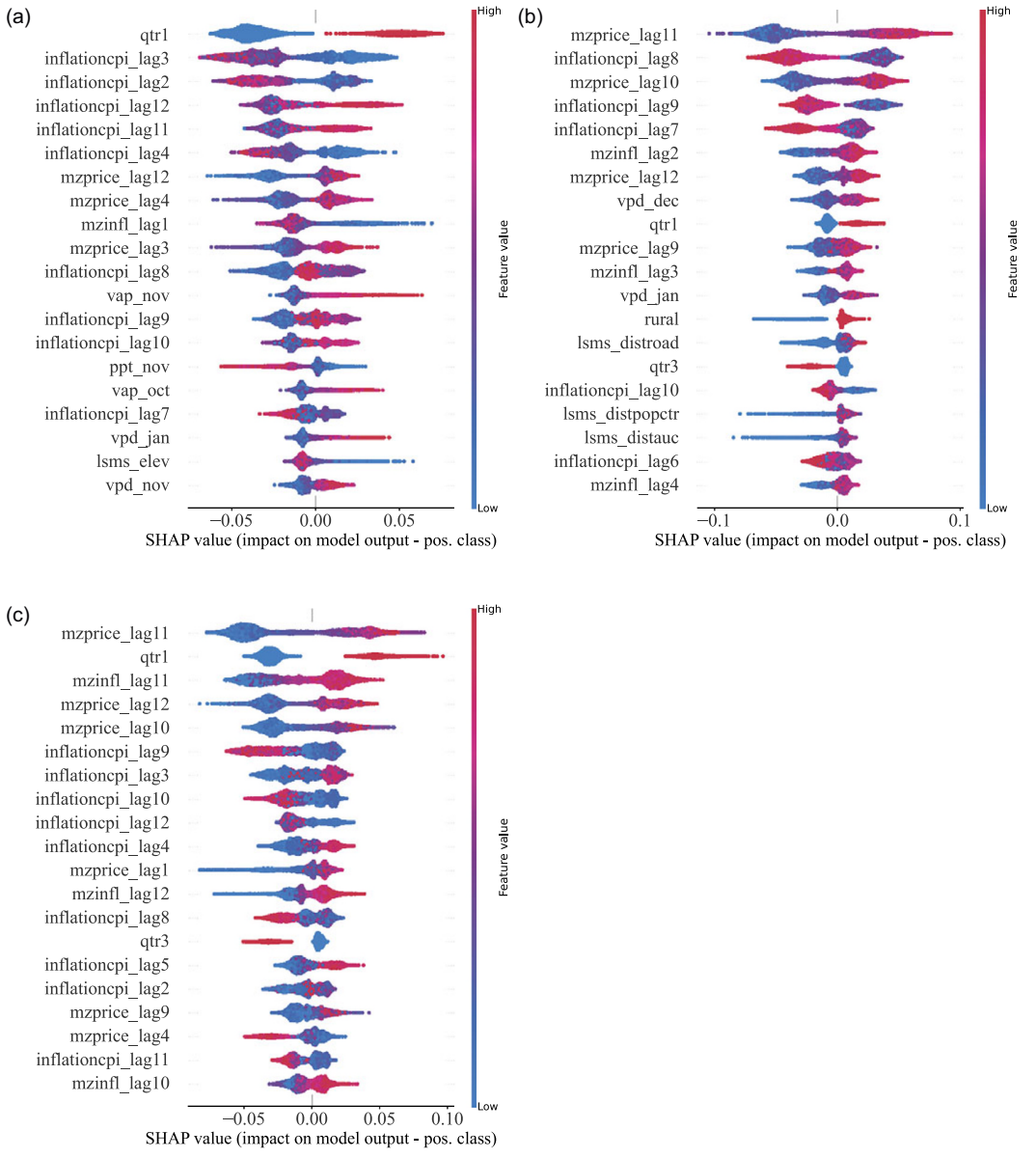

Figure 6. *SHAP decompositions for models trained and tested on data from wave 1 (top left), wave 3 (top right), and wave 4 (bottom),* c_1 *classification. Abbreviations: inflationcpi: monthly CPI inflation; lsms_elev: elevation; mzinfl: monthly maize price inflation; mzprice: maize price, ppt: precipitation, qtr: quarter; vap: vapor pressure; vpd: vapor pressure deficit. "Lag" indicates how far before the observation (in months) the value was taken.*

classification, the panel accuracy was 5.5 percentage points higher than the best-guess accuracy (82% and 76.5%), and at the c_2 classification, accuracy was roughly on par (90% versus 89%), although precision was substantially lower (0.3 versus 0.18). Comparisons across waves showed a positive trend in recall and a negative trend in precision; i.e., false positive assignments increase substantially while false negative assignments decline (Table 11).

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

Table 11. *Precision, recall, and accuracy of the nearest-neighbor matching (best-guess) approach for the c_1 and c_2 categories and each of the three waves for which predictions were generated*

Wave	Classification	Accuracy	Precision	Recall
2	c_1	77%	0.32	0.74
	c_2	89%	0.18	0.69
3	c_1	70%	0.49	0.74
	c_2	82%	0.36	0.56
4	c_1	72%	0.77	0.51
	c_2	83%	0.63	0.29

3.2.2. Predictions for the c_1 *classification*

At the c_1 classification, model fits initially underperformed the non-model predictor, although some achieved similar accuracy scores with high precision but low recall. Predictions improved as training data were added and occasionally surpassed the benchmark on accuracy in waves 3 and 4. In wave 4, the minimal, LSMS/IHS, and price specifications performed similarly, outperforming the benchmark by 7–10 percentage points in accuracy. The models frequently outperformed the precision benchmark, but recall was typically substantially lower. The only observed results surpassing both benchmarks occurred in wave 3 with logistic regression predicting on price and LSMS/IHS variables (Figure 7).

3.2.3. Predictions for the c_2 *classification*

Both the non-model classifier and ML modeled fits had the highest apparent accuracy in wave 2, but nearly half of the cross-validation runs (124 of 280) failed to produce any positive classifications, suggesting insufficient training data even with oversampling. The only variable sets producing consistent estimates were the LSMS variables and the remotely sensed variables, while price and inflation became more effective when predicting on waves 3 and 4. The non-modeled estimates gradually declined in accuracy across successive surveys, while ML model accuracy improved. In wave 4, RFs predicting from the remotely sensed variables or the price variables achieved accuracy rates of 90%, although overall precision and recall were low, and model fits were slightly less accurate than the minimal model. Recall was highest when using the LSMS/IHS variables and price variables, with LASSO regression achieving 0.87 recall on the former and 0.97 on the latter, with overall accuracies of 82 and 70%. Comparisons are presented in Figure 8.

The SHAP decompositions for the three sets of training data, when generated for all variables, initially ranked early and late season precipitation as significant contributors, but these variables were gradually displaced by prices and inflation. When the comparison was generated for solely the highest-performing inflation set, quarter fixed effects became most important, followed by maize price inflation lagged between 2 and 12 months (Figure 9).

4. Discussion

4.1. Evaluation of model performance and classification ability

Within-wave ML classifications did not consistently outperform logistic regression or a heuristic approach based on past information, but they did produce adequate fits. If judged on accuracy alone, the dataset was too "easy" (as defined by Grinsztajn et al., 2022), indicating that the additional work involved in ML may not be justified over a simpler method like logistic regression. However, we observed substantial variation in precision and recall, suggesting that marginal differences in accuracy could have substantial impacts when models are scaled; for example, in our wave four predictions on the c_2 classification, the RF predictions on the remotely sensed variables and price differed by less than a percentage point in accuracy (90.4% and 89.6%), but the remotely sensed specification made 30% more

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

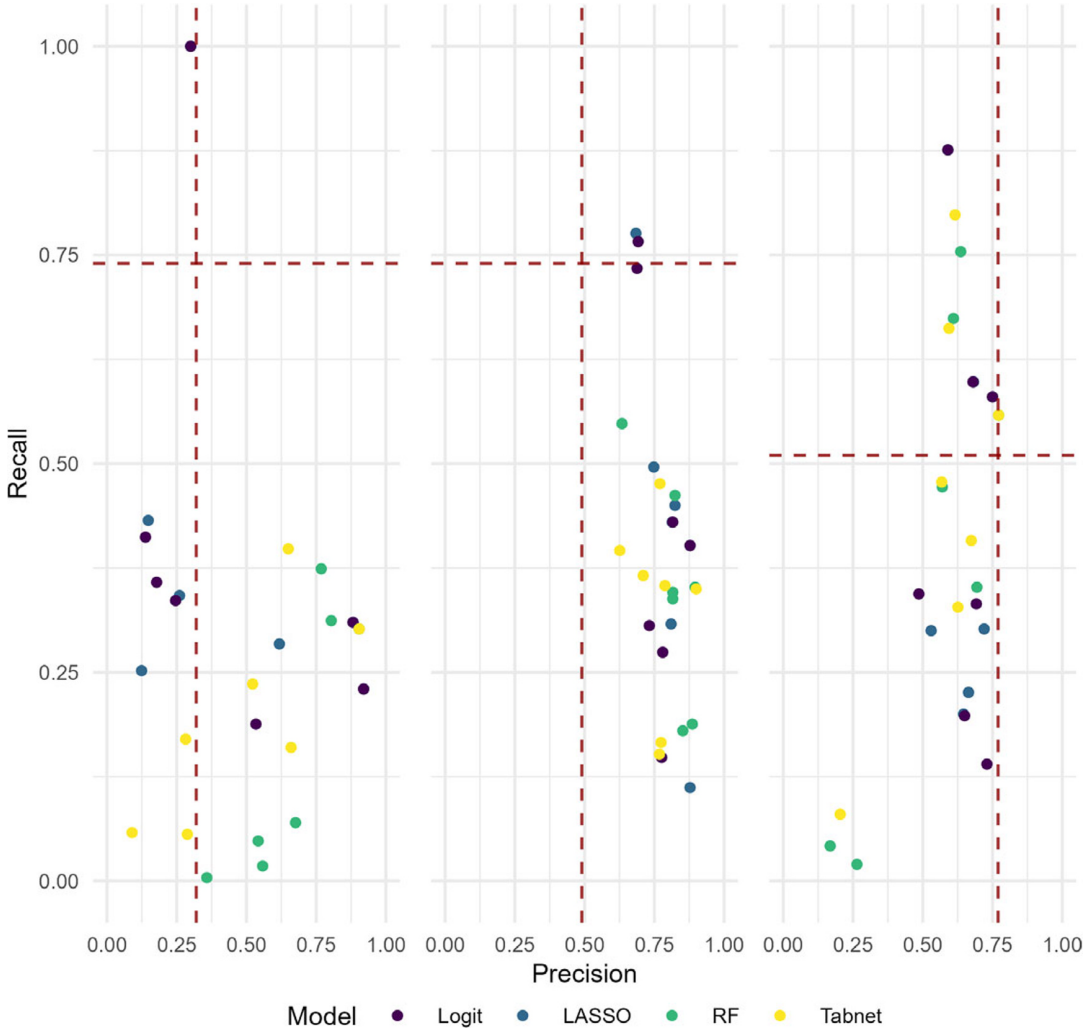

Figure 7. *Precision and recall scores across specifications and models on the* c_1 *classification compared to the benchmark precision and recall of the naïve classifier (red lines). The left plot represents predictions on wave 2 derived from wave 1, the center predictions on wave 3 derived from waves 1 and 2, and the right predictions on wave 4 derived from waves 1, 2, and 3.*

true positive assignments (192, compared to 148 for the price specification) while making a similar ratio of false positive assignments (283 for the remotely sensed specification and 223 for the price specification). Both underperformed on true positives but outperformed on false positives relative to the non-model method, with the non-model approach making 469 true positive assignments and 1,094 false-positive assignments. While the LSMS-ISA cannot be used to estimate EA-level populations because the household weights are designed to be nationally representative; estimates for the entire country were produced by the Netherlands Red Cross (2017) in their INFORM study. These estimates place the typical EA population at 1,600 individuals, with a maximum population of 16,081. Thus, even in this comparatively small-scale predictive exercise, sub-percentage-point margins in accuracy could generate differences of over 60,000 people targeted for aid, over half of whom may be affected by food insufficiency.

The difficulty in generating discrete, non-overlapping categorizations for EA-level food insufficiency at scales smaller than the resolution of our datasets combined with unaccounted for factors like assets and

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

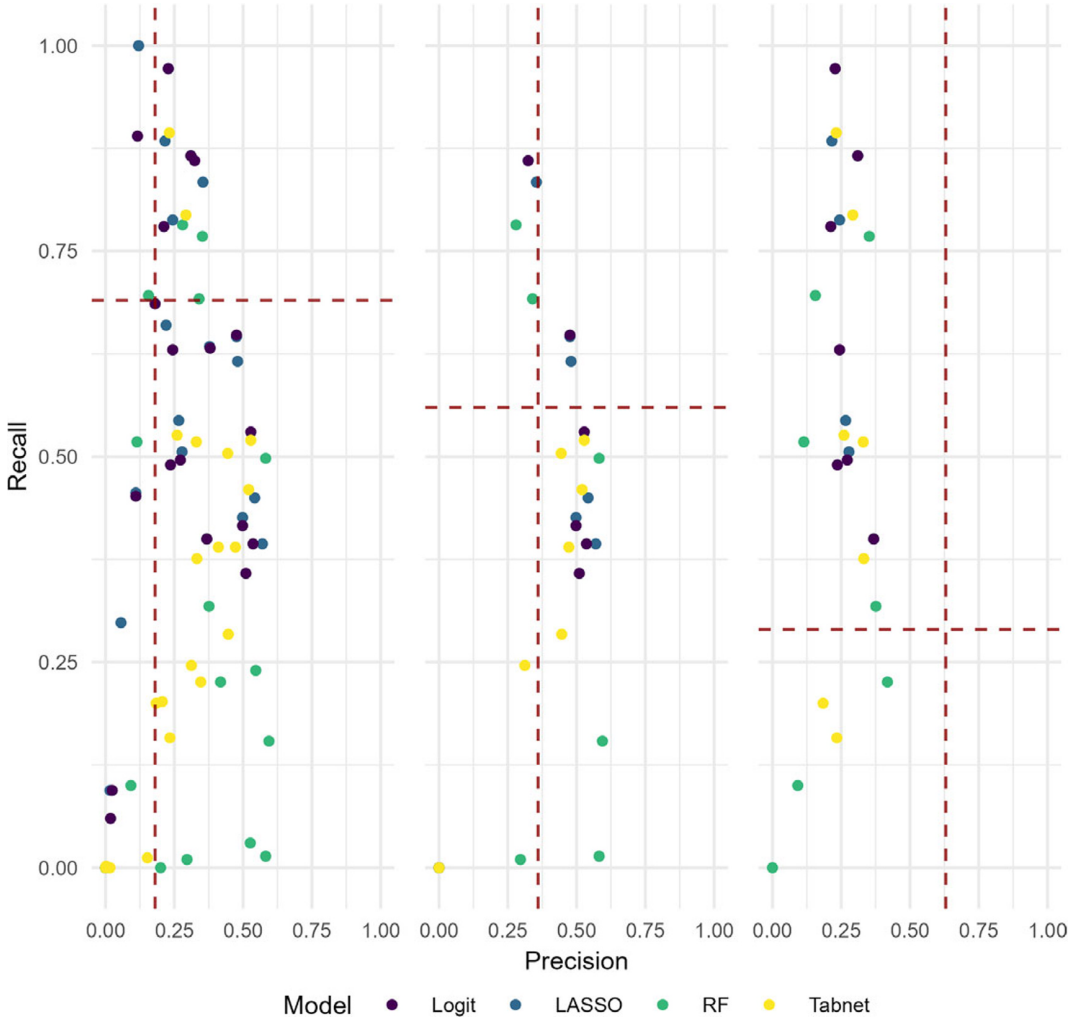

Figure 8. *Precision and recall scores across specifications and models on the* c_2 *classification compared to the benchmark precision and recall of the naïve classifier (red lines). The left plot represents predictions on wave 2 derived from wave 1, the center predictions on wave 3 derived from waves 1 and 2, and the right predictions on wave 4 derived from waves 1, 2, and 3.*

access to aid would have contributed to the higher false positive rate. Across specifications, the observed performance was highest on spatially interpolated price information and CPI inflation, suggesting that real-time weather observations may not add substantial information not already incorporated into prices, given the relative consistency of food insufficiency over time. Our findings on the limited value of ML in contexts with relatively strong spatial patterns over time are likely generalizable, but future research is necessary to both understand what those thresholds of weather patterns are and the value of ML in less predictable contexts.

Forecasts may provide advantages by giving more advanced notice of weather conditions, improving model lead times. At present, depending on the timing of rainfall, the Malawian government is typically aware if harvests will be poor by March (e.g., see WFP, 2024), which is already the peak of the current lean season. In previous crises, poor weather as early as October has presaged poor harvests (Ellis and Manda, 2012), but the depth of the advance warning will depend on the nature

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

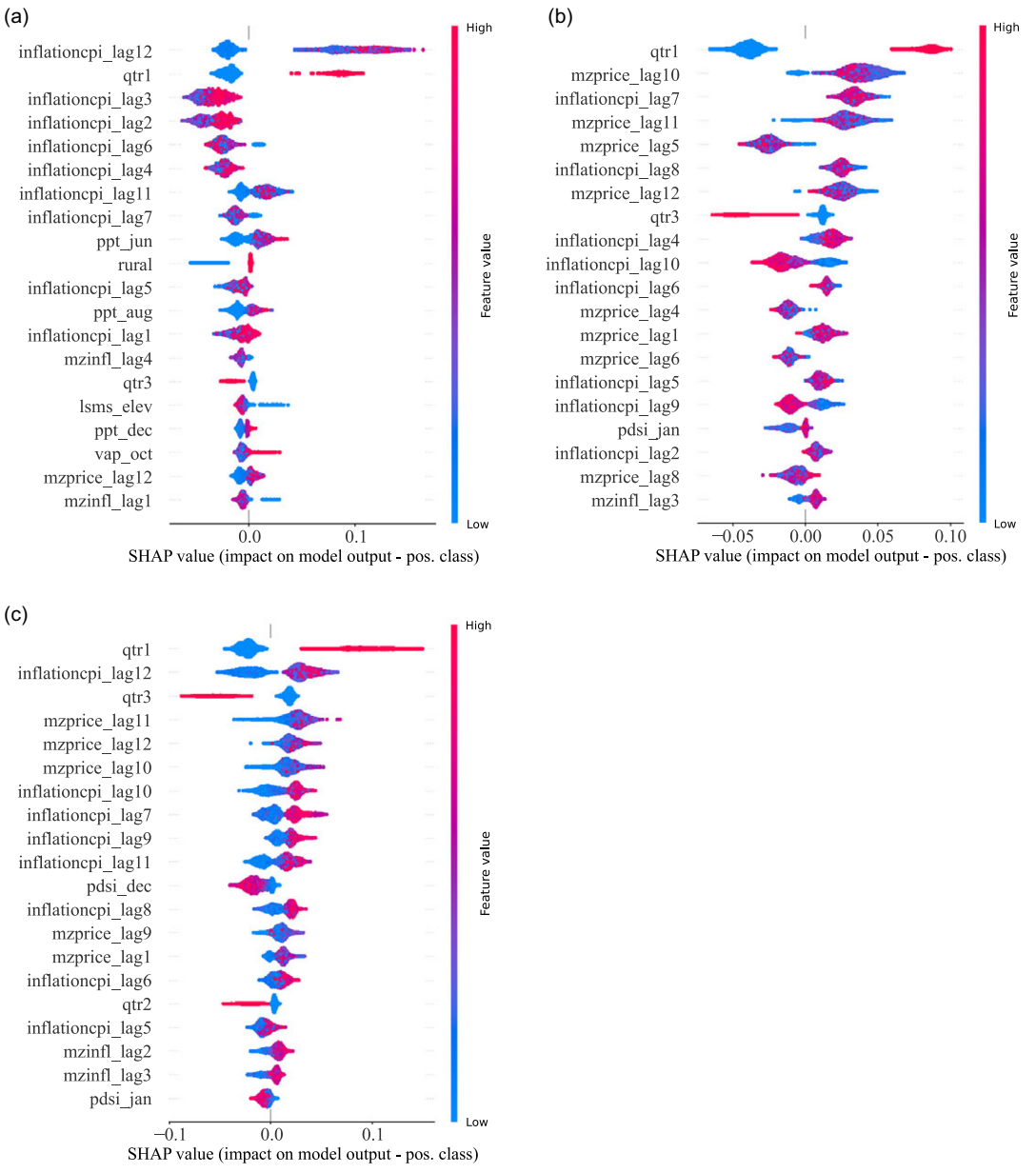

Figure 9. *SHAP decompositions for models trained and tested on data from wave 1 (top left), wave 3 (top right), and wave 4 (bottom),* c_1 *classification. Abbreviations: inflationcpi: monthly CPI inflation; mzinfl: monthly maize price inflation; mzprice: maize price, ppt: precipitation, tmin: mean daily minimum temperature; tmean: mean daily temperature; tmax: mean daily maximum temperature; qtr: quarter; vap: vapor pressure; vpd: vapor pressure deficit. "Lag" indicates how far before the observation (in months) the value was taken.*

and timing of the shock. This information will determine the extent and onset of the next lean season, whose severity will then be affected by off-season production, and the extent and delivery timeline of imported supplies (see Duchoslov et al., 2024). The former depends on residual soil moisture and the extent of irrigation, while the latter depends on policy and the international market. Delays in

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

securing imports further delays in their fulfilment, and insufficient aid for government purchasing have been associated with poor outcomes in the past (Ellis and Manda, 2012), and so advance notice may not provide information that decision-makers need or can act upon.

The relative success of the price and inflation variables along with their relatively high SHAP rank echoes findings by Martini et al. (2022), where food inflation was one of the strongest contributors to a RF model. These results indicate that sensitive predictions may be possible with a limited set of indicators, and food inflation may already capture information about the success or failure of the previous growing season, reducing the need for direct measurements of crop growing conditions, though further verification of consistency is warranted. Remote sensing also appeared to improve model fits in some scenarios, indicating that weather patterns within a given year may provide indications of the distribution of subnational variability, but the importance of these features fell as additional waves were added to generate forward predictions, suggesting that interannual variability in which weather conditions are leading to food insufficiency may limit the impact of these variables for long-term forecasts without expert data preparation or sustained training data collection.

4.2. Variable contributions

Based on the SHAP results, we find that multiple maize price indicators have the highest influence on model predictions. In Malawi, the relationship between hunger and price is complicated by households transitioning from net sellers in the post-harvest season to net buyers in the pre-harvest season (Cardell and Michelson, 2022). Because price increases harm net food purchasers but benefit net food sellers, Warr (2014) concludes that "the net effect of a change in food prices therefore depends on the sizes of these two groups and the amounts by which consumer and producer prices each change." The SHAP explanations for the RF models often produced double-headed plots in the price and inflation variables, where high values were found at either end of the axis rather than having a monotonic effect on predictions, a consequence of how the data were organized and the coverage of a single lagged observation. The effect of time as a model parameter was also visible when comparing models that predicted on observations only from the lean quarter to the predictions on the full dataset. The former produced lower accuracy scores even when compared to the full-year model's accuracy in the first quarter only.

Variables indicative of location, such as distances from market and elevation, appeared to be more effective at generating predictions for wave 2 observations compared to later waves or within individual waves, suggesting that there may have been a panel effect in sampling and reinforcing the need to test models over multiple time frames to verify continued validity. The remotely sensed variables also contributed more to fits within waves rather than across waves, suggesting that the models may be effective at noticing spatial variations over a single season but may encounter challenges integrating over a small number of cropping seasons if conditions are not consistent across them. In our data, we observed a consistent drought in Malawi's southern region, which contributed to food insufficiency in evaluations by the Famine Early Warning System, but periods of heavy rainfall from cyclones also contributed to nationwide food insufficiency in waves 3 and 4, something that didn't appear as a significant predictor in our models.

In general, what we refer to as the minimal approach consisting of solely dummies for time, region, and rural or urban residence, representing the prediction that a policymaker can make with minimal information, and the best-guess approach primarily relying on past instances of food insufficiency both produce results that would be considered reasonably accurate (i.e., around 80% classification accuracy). While these fixed effects showed significant stability during the study period, if significant changes occur —such as a region that typically experiences normal seasons suddenly facing a major drought or flood— short-term predictions may become substantially inaccurate over the next few periods; simultaneously, if updates occur in response to the event, the updated model may become less useful if the event does not recur for a long time.

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

4.3. Difficulties in accurately assessing food sufficiency

The data preparation process required overcoming both the challenges of defining the outcome appropriately and selecting the best variables from available data. Food insufficiency at the EA could have been structured as a categorical or continuous variable, but our preparatory work found that the models were better at making binary classifications rather than continuous or multinomially classified responses. While alternative indicators such as rCSI and HDDS that offer a more continuous form of measurement have been previously applied in food insufficiency modeling in Malawi (Lentz et al., 2019; Zhou, 2020 and Gholami et al., 2022), variations in what is being measured by different indices (see, e.g. Bertelli, 2019) and the availability of observations are considerations. In the LSMS IHS data, we observed high rates of spatiotemporal consistency and a lack of seasonal variation in HDDS. In the wave 1 and 2 surveys, we observed nearly complete overlap between the binned HDDS scores among the panel EAs (see Supplementary Material, Appendix C). The rCSI scores were considerably less consistent, with only 110 out of 204 EAs staying within bin between samples. Unlike the HDDS, rCSI varies seasonally and is thus sensitive to differences in interview dates across waves. These differences may explain why the HDDS modeled values were more accurate than the rCSI modeled values in Lentz et al. (2019). In a validation exercise, Vellema et al. (2015) found that individual consumption items provided little discrimination ability due to differences in dietary habits and that some food groups exhibited a negative relationship with dietary diversity. Thus, adding context-sensitive thresholding or splitting into a greater number of categories may improve the utility of HDDS and/or rCSI for ML-based models, but this may also require a greater number of observations for training.

In contrast to the LSMS/IHS methods of data collection, modeling efforts by Gholami et al. (2022) used high-frequency data with a sample concentrated in a smaller area where food insecurity was most widespread; RF models tend to perform better when the two labeled categories are closer to parity (Grinsztajn et al., 2022). Measurements of household assets or proxies of economic outlook also substantially contributed to model accuracy in this work, further supporting the conclusion that food insufficiency results more from challenges in distribution and social equity that may be difficult to sense remotely rather than biogeophysical factors influencing crop production. Two such factors identified in an analysis by Frelat et al. (2015) include off-farm income and livestock holdings, both of which represent alternative sources of income and/or direct provision of food for rural households and which account for approximately 20% and 12% of calories for food-insufficient households. A third factor is the ganyu labor system (a source of offseason agricultural employment for the rural poor): Sitienei et al. (2016) and Bouwman et al. (2021) both note how disruptions, such as introducing labor-saving herbicides, result in hunger due to lost employment opportunities (see also Fisher & Lewin, 2013). In addition to these factors, remote sensing crop productivity in Malawi appears to be more challenging than it is for its neighbors. Tang et al. (2022) found lower predictive accuracy in estimating crop yields in Malawi than in Tanzania. In addition, underperformance in Malawi and Ghana was observed by Gachoki and Muthoni (2023), and models for Malawi also underperformed models for Kenya in Lee et al. (2022). Li et al. (2022) achieved moderate accuracy but required higher-resolution satellite imagery.

5. Conclusion and implications for policy

While tools such as ML have the potential to detect warning signs of crises, for structural food insecurity whose geographic distribution may already be known and consistent or for crises that emerge suddenly from abrupt shifts in weather patterns, ML models are unlikely to provide information not already possessed by the experienced practitioner. Creating and maintaining models and troubleshooting problems that arise from changing standards or data sources is a challenge. Here, we demonstrate a simpler, easily explainable approach using existing data (naive matching) that produces predictions roughly on par with an ML model in some circumstances, but with tradeoffs in precision and recall that may be difficult for policymakers or aid organizations to navigate. Improvements in measurement and data collection frequency may change this assessment.

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

To explore whether usable forecasts can emerge from small sets of variables, we considered multiple types of publicly available and high- and low-frequency data: remotely sensed data on production indicators and price data, land cover, local topography, and distance to infrastructure. We found that these data can improve the predictive power of a simple model consisting solely of time and location but appear to capture similar levels of information about food sufficiency. The SHAP results indicate that price movements tend to offer more information about an EA's near-term food sufficiency status than remotely sensed factors related to crop productivity when the two are combined. Therefore, additional work on understanding the role of price as a leading indicator of food insufficiency and greater emphasis on collecting market observation data could lead to low-cost and rapidly effective forecasts to target areas for future intervention.

The variables particular to Malawi's food systems—persistent and reliable shortages by season, geographical consistency of food-insecure areas, and substantial reliance on a single staple crop—contribute to conditions for effective models trained on relatively small sets of variables. But they also lower the net value of any ML models for prediction and the generalizability of our finding non-modeled to contexts with more variable climate or dietary patterns. While continuing to build capacity to predict production will be useful, responding to food crises and working to prevent them appears to be the larger challenge, particularly in low-income countries with limited resources. Our findings contribute to an existing body of evidence (see Fadare, 2017) that policies focusing on smoothing seasonal food availability by supporting food staple affordability, dietary diversity and improving access to markets or postharvest storage regardless of anticipated production could provide significant benefits for food-insecure households and may reduce the need to rely on forecasts or predictions to ensure a stable food supply.

Abbreviations

EA	enumeration area
FCS	food consumption score
HDDS	household dietary diversity score
IHPS	Integrated Household Panel Survey
IHS	Integrated Household Survey
LSMS-ISA	Living Standards Measurement Study—Integrated Survey of Agriculture
ML	machine learning
rCSI	reduced coping strategies index
RF	random forests

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/dap.2025.10013.

Data availability statement. Data and code necessary to replicate the findings in this research are available at https://doi.org/10.5281/zenodo.15605219.

Acknowledgments. The authors are grateful to the three anonymous reviewers for their suggestions and Joaquin Mayorga for advice on data preparation.

Author contribution. SG: conceptualization, investigation, methodology, validation, visualization, and original draft; AT: data curation, investigation, methodology, validation, visualization, original draft, review and editing, and final draft; DA: conceptualization, investigation, project administration, supervision, review and editing, and final draft; CH: conceptualization and funding acquisition; DX: conceptualization; CB: conceptualization, review, and editing; RD: project administration; JL: project administration; CLA: conceptualization, project administration, resources, and review and editing.

Funding statement. This work was funded by the Microsoft AI for Good Laboratory and the Bill & Melinda Gates Foundation grant INV-043044.

Competing interest. The authors declare none.

References

Abatzoglou J, **Dobrowski S**, **Parks S and Hegewisch K** (2018) TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data 5*, 170191. https://doi.org/10.1038/sdata.2017.

Anderson CL, **Reynolds T**, **Merfeld JD and Biscaye P** (2017) Relating seasonal hunger and prevention and coping strategies: A panel analysis of Malawian farm households. *The Journal of Development Studies 54*(10), 1737–1755. https://doi.org/10.1080/00220388.2017.1371296.

Andrée BPJ (2021) *Estimating Food Price Inflation from Partial Surveys (Policy Research Working Paper 9886)*. World Bank Development Data Group

Arik SÖ and Pfister T (2021) TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence 35*(8). 6679–6687.

Ayalew M (1997) What is food security and famine and hunger? In Glantz, M.H. (ed), *Using Science against Famine: Food Security, Famine Early Warning, and El Niño*. Boulder, Colorado, USA: Cambridge University Press, pp. 1–8. https://www.ilankelman.org/glantz/Glantz1997ScienceFamine.pdf

Balashankar A, **Subramanian L and Fraiberger SP** (2023) Predicting food crises using news streams. *Science Advances 9*(9), eabm3449. https://doi.org/10.1126/sciadv.abm3449.

Barrett C (2010) Measuring food insecurity. *Science 327*(5967), 825–828. https://doi.org/10.1126/science.1182768.

Bergmeir C and Benitez JM (2012) On the use of cross-validation for time series predictor evaluation. *Information Sciences. 191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Bertelli O (2019) Food security measures in sub-Saharan Africa. A validation of the LSMS-ISA scale. *Journal of African Economies 29*(1), 90–120. https://doi.org/10.1093/jae/ejz011.

Bonuedi I, **Kornher L and Gerber N** (2022) Agricultural seasonality, market access, and food security in Sierra Leone. *Food Security 14*, 471–494.

Bouwman T, **Andersson J and Giller K** (2021) Herbicide induced hunger? Conservation agriculture, ganyu labour and rural poverty in Central Malawi. *The Journal of Development Studies 57*(2), 244–263. https://doi.org/10.1080/00220388.2020.1786062.

Cammack D (2012) Malawi in crisis, 2011-12. *Review of African Political Economy. 39*(132), 375–388.

Cardell L and Michelson H (2022) Price risk and small farmer maize storage in sub-Saharan Africa: New insights into a long-standing puzzle. *American Journal of Agricultural Economics 105*, 737–759. https://doi.org/10.1111/ajae.12343

Data4Diets: Building Blocks for Diet-related Food Security Analysis, Version 2.0. (2023). Tufts University, Boston, MA. Available at https://inddex.nutrition.tufts.edu/data4diets (accessed 24 February 2025).

Didan K (2015) *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 M SIN Grid V006. Data Set*. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD11A1.006

Duchoslav J, **Chiduwa M**, **Denhere S**, **De Weerdt J**, **Mzonde R and Phiri G** (2024) Responding to Malawi's impending food crisis. Available at <https://www.ifpri.org/blog/responding-malawis-looming-food-crisis/> (accessed 24 February 2025).

Ellis F and Manda E (2012) Seasonal food crises and policy responses: A narrative account of three food security crises in Malawi. *World Development 40*, 1407–1417. https://doi.org/10.1016/j.worlddev.2012.03.005.

Fadare O (2017) Effect of conflict and food Price shocks on calorie intake and acute malnutrition in Nigeria: A micro-panel data analysis. In *93rd Annual Conference of the Agricultural Economics Society*, University of Warwick, England 15–17 April 2019.

FAO, IFAD, UNICEF, WFP and WHO (2024) *The State of Food Security and Nutrition in the World 2024—Financing to End Hunger, Food Insecurity and Malnutrition in all its Forms*. Rome

Fick S and Hijmans RJ (2017) WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology 37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fisher M and Lewin P (2013) Household, community, and policy determinants of food insecurity in rural Malawi. *Development Southern Africa 30*(4–5), 451–467.

Frelat R, **Lopez-Ridaura S**, **Giller KE**, **Herrero M**, **Douxchamps S**, **Djurfeldt AA**, **Erenstein O**, **Henderson B**, **Kassie M**, **Paul BK**, **Rigolot C**, **Ritzema RS**, **Rodriguez D**, **van Asten PJA and van Wijk MT** (2015) Drivers of household food availability in sub-Saharan Africa based on big data from small farms. *Proceedings of the National Academy of Sciences 113*(2), 458–463. https://doi.org/10.1073/pnas.1518384112.

Friedl M and Sulla-Menashe D (2019) *MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006* [Data Set] (Tech. Rep.). NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD12Q1.006

Gachoki S and Muthoni F (2023) Drivers of maize yield variability at household level in northern Ghana and Malawi. *Geocarto International 38*, 1. https://doi.org/10.1080/10106049.2023.2230948.

Gebrehiwot T and Van der Veen A (2014) Coping with food insecurity on a micro-scale: Evidence from Ethiopian rural households. *Ecology of Food and Nutrition 53*(2), 214–240. https://doi.org/10.1080/03670244.2013.811387.

Gholami S, **Knippenberg E**, **Campbell J**, **Andriantsimba D**, **Kamle A**, **Parthasarathy P**, **Sankar R**, **Birge C and Lavista Ferres J** (2022) Food security analysis and forecasting: A machine learning case study in southern Malawi. *Data & Policy 4*, e33. https://doi.org/10.1017/dap.2022.25.

Grinsztajn L, **Oyallon E and Varoquaux G** (2022) *Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?* NeurIPS 2022 Datasets and Benchmarks Track.

Khandker S and Mahmud W (2012) *Seasonal Hunger and Public Policies: Evidence from Northwest Bangladesh*. Washington, DC: World Bank. https://doi.org/10.1596/978-0-8213-9553-0

Krishnamurthy RP, **Fisher J**, **Choularton R and Kareiva P** (2022) Anticipating drought-related food security changes. *Nature Sustainability 5*, 956–964.

https://doi.org/10.1017/dap.2025.10013 Published online by Cambridge University Press

Lee D, **Davenport F**, **Shukla S**, **Husak G**, **Funk C**, **Harrison L**, **McNally A**, **Rowland J**, **Budde M and Verdin J** (2022) Maize yield forecasts for sub-Saharan Africa using earth observation data and machine learning. *Global Food Security 33*(100643). https://doi.org/10.1016/j.gfs.2022.100643.

Lentz E, **Michelson H**, **Baylis K and Zhou Y** (2019) A data-driven approach improves food insecurity crisis prediction. *World Development 122*, 399–409. https://doi.org/10.1016/j.worlddev.2019.06.008.

Li C, **Chimimba EG**, **Kambombe O**, **Brown LA**, **Chibarabada TP**, **Lu Y**, **Anghileri D**, **Ngongondo C**, **Sheffield J and Dash J** (2022) Maize Yield Estimation in Intercropped Smallholder Fields Using Satellite Data in Southern Malawi. *Remote Sensing 14* (10), 2458. https://doi.org/10.3390/rs14102458

Madsden S, **Bezner Kerr R**, **LaDue N**, **Luginaa I**, **Dzanja C**, **Dakishoni L**, **Lupafya E**, **Shumba L and Hickey C** (2021) Explaning the impact of agroecology on farm-level transitions to food security in Malawi. *Food Security 13*, 933–954. https://doi.org/10.1007/s12571-021-01165-9.

Martini G, **Bracci A**, **Riches L**, **Jaiswal S**, **Corea M**, **Rivers J**, **Husain A and Omodei E** (2022) Machine learning can guide food security efforts when primary data are not available. *Nature Food 3*(9), 716–728. https://doi.org/10.1038/s43016-022-00587-8.

Maxwell D and Caldwell R (2008) *The Coping Strategies Index: Field Methods Manual—Second Edition*. Cooperative for Assistance and Relief Everywhere, Inc. (CARE).

Maxwell D, **Khalif A**, **Hailey P and Checchi F** (2020) Viewpoint: Determining famine: Multi-dimensional analysis for the twenty-first century. *Food Policy 92*, 101832.

McCarthy N, **Kilic T**, **de la Fuente A and Brubaker J** (2018) Shelter from the storm? Household-level impacts of, and responses to, the 2015 floods in Malawi. *Economics of Disasters and Climate Change 2*, 237–258. https://doi.org/10.1007/s41885-018-0030-9.

Netherlands Red Cross (2017) Malawi—INFORM-based prioritization of enumeration areas. Dataset. Available at <https://data.humdata.org/dataset/inform-based-prioritization-of-enumeration-areas-in-malawi>

Sitienei I, **Mishra A and Khanal A** (2016) Informal "Ganyu" labor supply, and food security: The case of Malawi. *Food Security in a Food Abundant World 16*, 159–175. https://doi.org/10.1108/S1574-871520150000016015.

Tang B, **Liu Y and Matteson DS** (2022) Predicting poverty with vegetation index. *Applied Economic Perspectives and Policy 44*(2), 930–945.

Vaitla B, **Devereux S and Swan SH** (2009) Seasonal hunger: A neglected problem with proven solutions. *PLoS Medicine 6*(6), e1000101. https://doi.org/10.1371/journal.pmed.1000101.

Vellema W, **Desiere S and D'Haese M** (2015) Verifying validity of the household dietary diversity score: An application of Rasch Modeling. *Food and Nutrition Bulletin 37*, 1. https://doi.org/10.1177/0379572115620966.

Villacis A, **Badruddoza S**, **Mishra A and Mayorga J** (2023) The role of recall periods when predicting food insecurity: A machine learning application in Nigeria. *Global Food Security 36*, 100671. https://doi.org/10.1016/j.gfs.2023.100671.

Wan Z, **Hook S and Hulley G** (2015) *MOD110A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006* [Data Set] (Tech. Rep.). NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD11A1.006.

Warr P (2014) Food insecurity and its determinants. *Australian Journal of Agricultural and Resource Economics 58*, 519–537. https://doi.org/10.1111/1467-8489.12073.

Westerveld JJ, **van den Homberg MJ**, **Nobre GG**, **van den Berg DL**, **Teklesadik AD and Stuit SM** (2021) Forecasting transitions in the state of food security with machine learning using transferable features. *Science of the Total Environment 786*, 147366. https://doi.org/10.1016/j.scitotenv.2021.147366.

WFP. Urgent Action Critical as Malawi Faces Severe Drought. (2024, May 14). https://www.wfp.org/news/urgent-action-critical-malawi-faces-severe-drought

Zhou Y (2020) *Three Essays on Machine Learning and Food Security*. Doctoral dissertation, University of Illinois at Urbana Champaign.

Zhou Y, **Lentz E**, **Michelson H**, **Kim C and Baylis K** (2022) Machine learning for food insecurity: Principles for transparency and usability. *Applied Economic Perspectives and Policy 44*, 893–910. https://doi.org/10.1002/aepp.13214.

Cite this article: Tomes A, Gholami S, Alia D, Hennessy C, Xu D, Bitz C, Dodhia R, Lavista Ferres J and Anderson CL (2025). Assessing the utility of machine learning for predicting food sufficiency: a case study in Malawi. *Data & Policy*, 7: e52. doi:10.1017/dap.2025.10013