npj | clean water

www.nature.com/npjcleanwater

ARTICLE OPEN

Addressing gaps in data on drinking water quality through data integration and machine learning: evidence from Ethiopia

Alemayehu A. Ambel[1]✉, Robert Bain[2], Tefera Bekele Degefu[1], Ayca Donmez[2], Richard Johnston[3] and Tom Slaymaker[2]

Monitoring access to safely managed drinking water services requires information on water quality. An increasing number of countries have integrated water quality testing in household surveys however it is not anticipated that such tests will be included in all future surveys. Using water testing data from the 2016 Ethiopia Socio-Economic Survey (ESS) we developed predictive models to identify households using contaminated (≥1 *E. coli* per 100 mL) drinking water sources based on common machine learning classification algorithms. These models were then applied to the 2013–2014 and 2018–2019 waves of the ESS that did not include water testing. The highest performing model achieved good accuracy (88.5%; 95% CI 86.3%, 90.6%) and discrimination (AUC 0.91; 95% CI 0.89, 0.94). The use of demographic, socioeconomic, and geospatial variables provided comparable results to that of the full features model whereas a model based exclusively on water source type performed poorly. Drinking water quality at the point of collection can be predicted from demographic, socioeconomic, and geospatial variables that are often available in household surveys.

npj Clean Water (2023)6:63 ; https://doi.org/10.1038/s41545-023-00272-8

INTRODUCTION

In many low- and middle-income countries the majority of the population lacks access to drinking water services that are free from contamination, accessible on premises and available when needed. In 2020, around 26% of the global population (2 billion people) lacked safely managed drinking water, rising to 71% in low-income countries[1]. Drinking water quality, specifically fecal contamination of drinking water sources, is often found to be the limiting factor for safely managed drinking water services[2]. Contaminated drinking water can transmit diseases such as diarrhea, cholera, dysentery, typhoid, and polio and lack of safe water is associated with substantial disease burden. An estimated 829,000 diarrheal deaths are attributed to inadequate drinking water, sanitation and hygiene each year[3]. Expanding access to safe drinking water is thus an important human development priority with targets set at national and global levels. The Sustainable Development Goals (SDG) target 6.1 calls for “safe” drinking water “for all” by 2030 and the associated safely managed drinking water services indicator requires information on the quality of drinking water[4].

In the absence of robust water quality data from administrative systems, an increasing number of National Statistical Offices (NSOs) in low- and middle-income countries have integrated water quality testing in nationally representative household surveys to generate data on water quality and baselines for safely managed drinking water services[2]. In these countries, field teams have tested 100 mL of drinking water for WHO’s preferred indicator of fecal contamination in drinking water, *Escherichia coli* (*E. coli*). To meet WHO guidelines drinking water should not contain any detectable *E. coli* in any 100 mL sample[5]. Key advantages of this approach are ensuring data are representative and cover the entire population, including households reliant on informal services or self-supply, and the ability to link water quality information to the wealth of other information collected in household surveys. The integration of *E. coli* testing in household surveys is, however, an additional burden on field teams and requires equipment and consumables as well as dedicated training in aseptic techniques, incubation and interpreting results. Therefore, it is not expected that the module will be included in all future surveys conducted by NSOs, but rather that the module might be repeated every 3–5 years.

Machine learning techniques hold promise for predicting water quality using data from household surveys given the wealth of information on household socio-economic conditions and a range of geospatial information from global datasets that can be integrated using the household survey cluster locations (where available). Prior research drawing on data from the World Bank’s Living Standard Measurement Study (LSMS) surveys has applied machine learning to a wide range of topics covered in these surveys including poverty[6], housing rental value[7], food security[8], crop type mapping[9], crop yield[10], and fertilizer pricing[11]. Recent studies have investigated the ability to predict microbial drinking water quality on smaller scales for specific water source types including piped water in the Democratic Republic of Congo[12] and groundwater in Uganda and Bangladesh[13]. Studies have also utilized machine learning algorithms to generate national and global predictive maps for arsenic[14] and fluoride[15] and to develop predictive models for microbial contamination of surface[16] and recreational waters[17]. To our knowledge machine learning has not previously been applied to drinking water quality data from a nationally representative sample of households nor has it been used to predict water quality for surveys that did not include direct measures of water quality.

Here we examine the performance of a range of commonly used algorithms to predict *E. coli* contamination in drinking water sources in Ethiopia. This study draws on the results of the third wave of the Ethiopia Socioeconomic Survey (ESS3) in 2016 to examine the performance of machine learning algorithms in predicting water quality. The objectives of this study were: (I) to predict contamination of drinking water sources and assess the

[1]Development Data Group, World Bank, Washington, DC, USA. [2]Division of Data, Analysis, Planning and Monitoring, UNICEF, New York, NY, USA. [3]Department of Environment, Climate Change and Health, WHO, Geneva, Switzerland. ✉email: aambel@worldbank.org

relative performance of commonly used machine learning algorithms (II) to examine the role of different groups of variables (household characteristics, water service characteristics, geospatial variables) on predictions and (III) to apply the highest performing predictive model to the other waves of the ESS in 2013–2014 (ESS2) and 2018–2019 (ESS4) which did not incorporate water quality testing.

RESULTS

Drinking water quality

According to the 2015/16 ESS, nationally, about 68% of households fetched their drinking water from improved sources such as piped water, bottled water, protected springs, and wells (Table 1)[18]. This is the same level (68%) as reported in the 2016 Ethiopia Demographic and Health Survey (DHS)[19]. About a third of households collected their water from unprotected sources such as springs, lakes, and ponds. Despite about two-thirds of the population using improved sources, water testing in the ESS demonstrated that many of these sources are not free from *E. coli* contamination (Table 1). *E. coli* contamination was detected in 84% (95% confidence interval (CI) 82%, 87%) of the households' water sources. There were only 864 households (15.6%) with water sources free of *E. coli* contamination. As has been found in previous studies[20], contamination rates are in general lower in samples from improved sources (piped sources and protected springs and wells) than from unimproved sources (unprotected springs and surface water). Yet, over 78% (95% CI 74%, 81%) of the samples collected from households who fetched their drinking water from improved sources are contaminated. The contamination rate varies by improved source type; for example, among users of piped water, contamination rates range from 58% for piped water on-premises to 74% for public taps. Improved sources with the highest contamination rates are protected springs and wells; *E. coli* was detected from over 89% of the households who reported using these sources. The detection of contamination in all types of improved sources confirms that improved sources are not necessarily safe sources and points to the fact that the commonly used practice of assessing access to drinking water services based on the type of water source ("improved" vs "unimproved) provides an incomplete picture in countries like Ethiopia[21].

Predicting water quality from socioeconomic information

Classification algorithms. Table 2 presents model selection results using six classification algorithms including Extreme Gradient Boosting (XGBoost), Generalized Linear Model (GLM), Generalized linear models with elastic net regularization (GLMNET), K-nearest neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM). The RF algorithm produced the best performance across all metrics. It produced the highest discrimination ability with the area under the receiving operator curve (AUC) of 0.91 (95% CI 0.89, 0.94). This model is also superior in other performance metrics including accuracy, F1 score, sensitivity,

Table 1. *E. coli* contamination of drinking water at the point of collection by source type.

Drinking water source types	Number of households	Proportion of households with the source	Proportion of *E. coli* contamination at the point of collection
A. Improved Sources			
Piped on-premises	1143	0.19	0.58 (0.51,0.65)
Piped water public	507	0.12	0.74 (0.62,0.87)
Tanker/Vendor	219	0.04	0.71 (0.57,0.85)
Protected springs, wells, boreholes	1289	0.32	0.90 (0.87,0.94)
Rainwater	47	0.01	0.97 (0.92,1.02)
Improved (all sources)	3205	0.68	0.78 (0.74,0.81)
B. Unimproved Sources			
Unprotected springs, wells	897	0.22	0.98 (0.95,1.00)
Surface water	498	0.09	1.00 (0.99,1.00)
Other	88	0.01	0.82 (0.66,0.97)
Unimproved (all sources)	1483	0.32	0.98 (0.96, 1.00)
All Sources	4688	1.00	0.84 (0.82,0.87)

The household numbers are unweighted, proportions are weighted, and values in the parenthesis in the last column are 95% confidence intervals.

Table 2. Comparison of classification algorithms.

Algorithm	Accuracy (95%CI)	F1-score	Sensitivity	Specificity	AUC (95%CI)
RF	0.89 (0.87, 0.91)	0.93	0.95	0.64	0.91 (0.89, 0.94)
XGBoost	0.88 (0.85, 0.90)	0.92	0.94	0.62	0.90 (0.88, 0.93)
SVM	0.83 (0.81, 0.86)	0.90	0.92	0.46	0.82 (0.78, 0.86)
GLMNET	0.85 (0.82, 0.87)	0.91	0.95	0.43	0.85 (0.82, 0.88)
GLM	0.85 (0.82, 0.87)	0.91	0.95	0.42	0.85 (0.82, 0.88)
KNN	0.84 (0.82, 0.87)	0.91	0.93	0.49	0.85 (0.82, 0.88)

Results are based on the 2015/16 Ethiopia Socioeconomic Survey data. Accuracy results are significantly higher than the no information rate (NIR) of 0.80 in all models. See Supplementary Table 2 for more results.

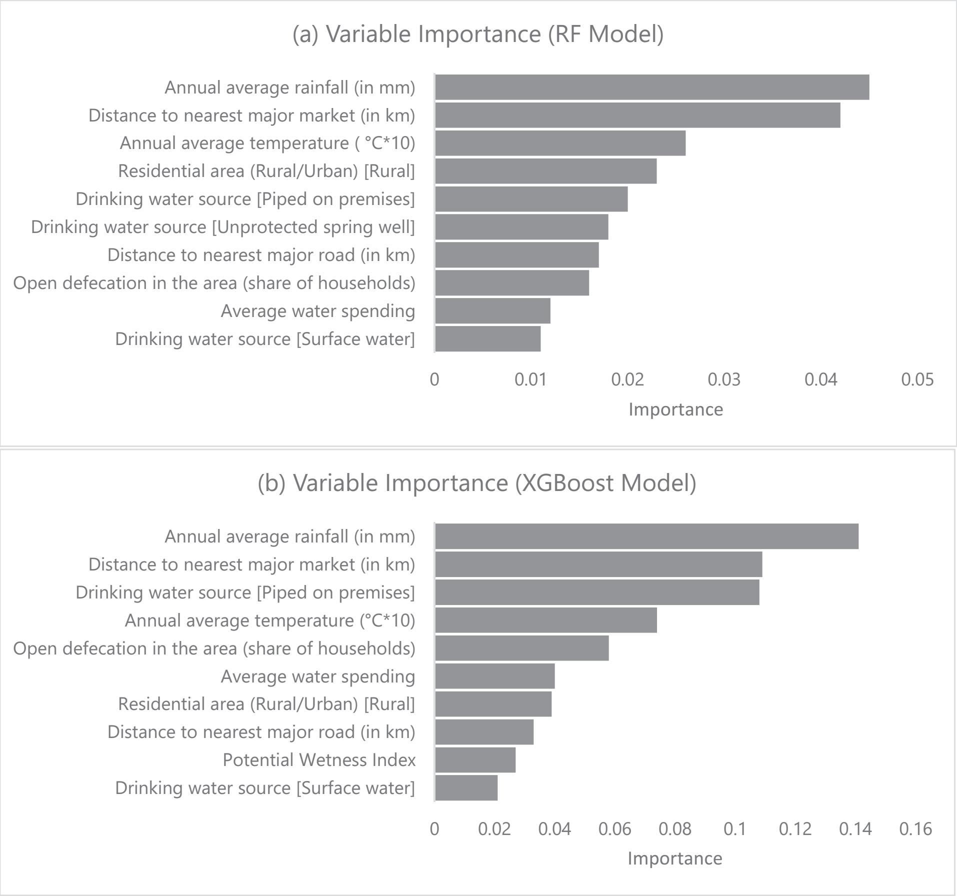

Fig. 1 Variable Importance. Variable importance for RF and XGBoost models - top 10 features.

and specificity. Another ensemble method, XGBoost provided a comparable performance with an AUC of 0.90 (95% CI 0.88, 0.93) and high performance in other metrics. The remaining four algorithms, GLM, GLMNET, SVM, and KNN show high predictive performance but the results in these models are lower than the RF model.

Variable importance. Figure 1 presents the top 10 predictors which are found to be quite similar across the RF and XGBoost models (See Supplementary Tables 7 and 8 for top 20 predictors). These important features in both models are constructed from a few commonly included questions in household surveys. These are geographic variables, water source type, location of the household, housing characteristics, and assets. Out of the top 10 features that are strongly associated with *E. coli* contamination, 50% in the RF and 60% in the XGBoost model are geospatial variables. These variables include region, place of residence (rural or urban), distance to a major road or a major market center, average rainfall, wetness index, and temperature of the area where the household is located. The second most important features relate to household characteristics including the type of water source used by household members, the proportion of households without a toilet facility in the area (open defecation) and the amount households spend on drinking water.

Model performance using different sets of features. The variable importance figures for the full model show that features related to geospatial variables and water source types appear to be strong predictors of contamination of drinking water sources. We now compare the performance of different scenarios with the objective of identifying models that can predict contamination with minimal socioeconomic information. We examined models with the following features: (i) water source type only, (ii) water source and selected household-level variables but excluding geospatial variables, (iii) geospatial variables only, and (iv) geospatial and household level variables but excluding water source types. These scenarios were selected based on their availability in a range of household surveys and other data collection activities. For example, water source type is an important indicator often included in household surveys and is used to measure access to improved sources of drinking water[22]. Similarly, recent surveys often capture GPS coordinates of the sampled households or the center of the census enumeration area. GPS coordinates can be used to generate several geospatial variables without extra burden to the survey but not all surveys currently make these coordinates available to researchers. Other household-level characteristics and basic demographic information about household members are often included in LSMS or DHS type surveys but may not be available in other assessments, such as Ethiopia’s national WASH inventory[23].

Table 3 presents results of different feature scenarios for the RF model. As noted earlier, the full model that uses all the features has a strong discrimination ability (AUC 0.91; 95% CI 0.89, 0.94). The rest of the models have lower discrimination ability with the least performing being the model when water source type is considered as the only predictor (AUC 0.80; 95% CI 0.77, 0.84). However, augmenting water source variables with selected household-level variables, but excluding geospatial variables, resulted in a performance comparable to the full model (AUC 0.89; 95% CI 0.86, 0.91). The geospatial features only model has a performance (AUC 0.91; 95% CI 0.88, 0.93) which is also comparable with the full model. The performance did not

Table 3. Prediction performance results in different model scenarios.

Model Scenario	Accuracy (95%CI)	F1-score	Sensitivity	Specificity	AUC (95%CI)
All Features	0.89 (0.87, 0.91)	0.93	0.95	0.64	0.91 (0.89, 0.94)
Water Source Only	0.80 (0.80, 0.80)	0.89	1.00	0.00	0.80 (0.77, 0.84)
Water Source & Household Variables	0.85 (0.83, 0.87)	0.91	0.95	0.46	0.89 (0.86, 0.91)
Geospatial Only	0.88 (0.85, 0.90)	0.92	0.95	0.58	0.91 (0.88, 0.93)
Geospatial & Household Variables	0.87 (0.85, 0.89)	0.92	0.95	0.56	0.90 (0.87, 0.93)

Results are based on the 2015/16 ESS data. Accuracy results are significantly higher than the no information rate (NIR) of 0.80 in all but the "Water Source Only" scenario where the model predicted all water sources to be contaminated. See Supplementary Tables 5 and 6 for both train and test data results from RF and XGboost models.

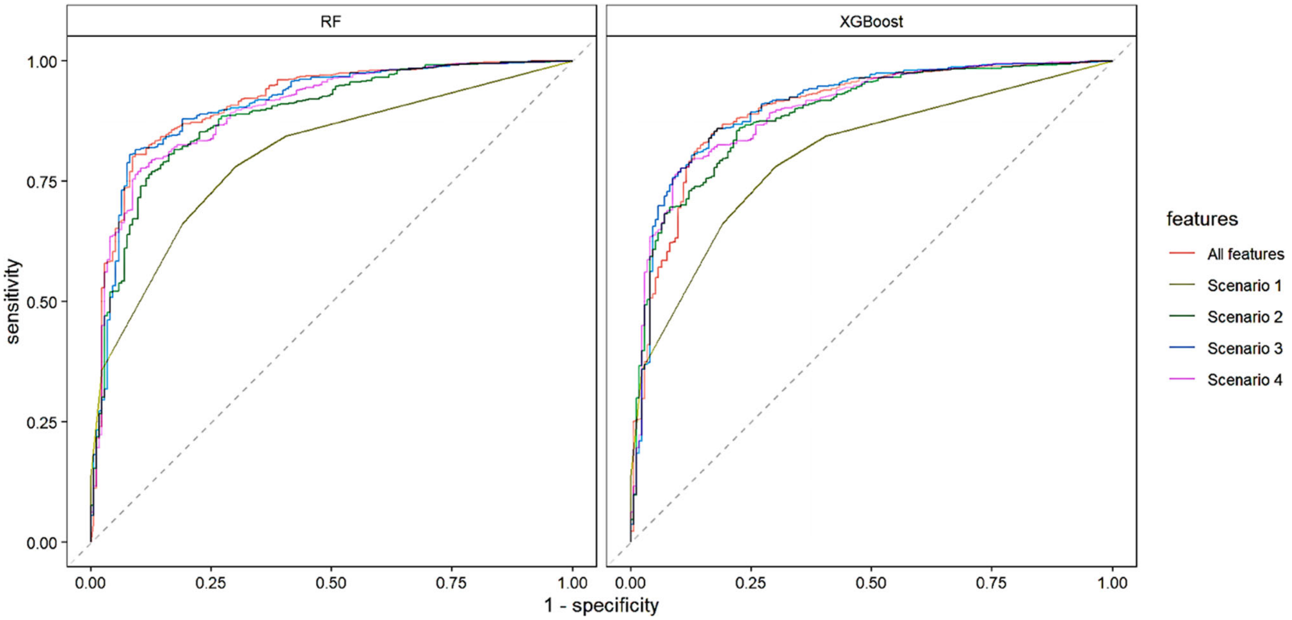

Fig. 2 ROC curves for model scenarios. Receiver operating characteristic (ROC) curves for model scenarios using RF and XGBoost classification algorithms.

however improve when selected household characteristics were added to the geospatial features model (AUC 0.90; 95%CI 0.87, 0.89). This is also illustrated in the ROC curves for the RF and XGBoost models, i.e., the ROC curve for the water source only model is the shallowest in both cases (Fig. 2).

Classification results. Table 4 presents the distribution of classification results for the full features RF model by residence and water source type (Supplementary Table 5 includes performance metrics). Overall, the predictive model classified 88.4% of the cases correctly. The performance is higher in rural (93.6%) than urban (77.8%) areas. There was some variation in classification performance by water source type, with at least 79.4% of cases correctly classified except for truck/vendor where performance dropped to 61%.

Table 4 also presents classification results for the four scenarios. The water source-only model predicted that all water sources were contaminated and thus correctly classified all the contaminated sources but misclassified all the non-contaminated sources as contaminated. Addition of selected household variables in the Water Source Plus model, greatly improved specificity and the classification results became comparable with the full model. The geospatial features only model performed well and provided a comparable performance with that of the full model (88.1%).

Model application in different datasets. Lastly, we applied the predictive model to two other waves of the ESS fielded in 2013/14 and in 2018/19 which did not include a water quality testing module. These two surveys have information on all the socio-economic and geospatial features identified in the predictive model using the 2015/16 survey data. Whereas the 2013/14 and the 2015/16 survey waves interviewed the same households, the sample was refreshed for the 2018/19 wave of the ESS and thus fielded in different households.

Figures 3 and 4 compare the actual proportions of contaminated and non-contaminated water sources in 2015/16 against the predicted values in all three survey waves. The results are for the full model and the scenarios (Fig. 3) and water source types (Fig. 4). As expected, the actual and predicted values are very close for 2015/16. The comparison of prediction results across survey waves shows slight differences in the households' access to drinking water free from *E. coli*; access was better in the reference survey than in the preceding and subsequent rounds of the ESS. This pattern is reflected in all models.

Disaggregating predictions by water source type shows that the results are comparable for most source type categories (Fig. 4). The actual in 2015/16 and the predicted values in the three survey waves are close or within the 95% confidence interval for the following water source types: piped on-premises, rainwater, unprotected springs and wells, and surface water. The differences

Table 4. Classification results from RF prediction model by residence, water source type, and model scenario.

	Number of households	Correctly Classified		Misclassified		Total Correctly Classified (%)
		Contaminated (%)	Not Contaminated (%)	Contaminated (%)	Not Contaminated (%)	
a. Full Model						
National	883	76.3	12.1	7.5	4.1	88.4
Urban	288	47.2	30.6	12.2	10.1	77.8
Rural	595	90.4	3.2	5.2	1.2	93.6
Piped on-premises	214	41.1	38.3	9.8	10.7	79.4
Public standpipe	103	77.7	13.6	4.9	3.9	91.3
Truck, vendor	41	56.1	4.9	26.8	12.2	61.0
Rainwater	6	100.0	0.0	0.0	0.0	100.0
Protected spring/well	245	86.5	2.4	9.4	1.6	89.0
Unprotected springs, well	159	97.5	0.6	1.9	0.0	98.1
Surface water	99	100.0	0.0	0.0	0.0	100.0
b. Scenarios						
Water Source only	883	80.4	0.0	19.6	0.0	80.4
Water Source Plus	883	76.2	9.4	10.2	4.2	85.6
Geospatial only	883	76.7	11.4	8.2	3.7	88.1
Geospatial Plus	883	76.1	10.6	8.9	4.3	86.7

The total number of households included in this study is 4688. The total number of households in column 2 of this table refers to the number of households in the test data.

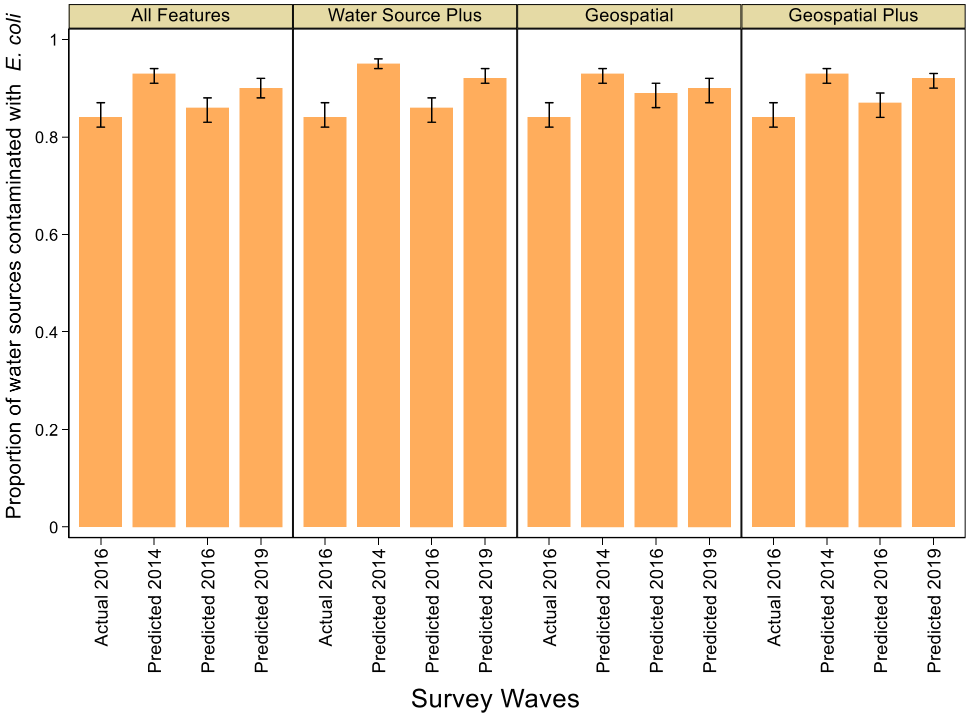

Fig. 3 Prediction results by model scenario. Prediction results of *E. coli* contamination of drinking water sources in three survey waves by model scenario. Error bars represent 95% confidence intervals.

across survey waves are concentrated in two categories of improved sources namely public standpipes and water tanker and water kiosk/vendors. Overall, the prediction results in Figs. 3 and 4 show that the general pattern is maintained, i.e., the average contamination rate is always close to 90% and by source type, it is the highest in the unimproved source categories and the lowest in the pipes on-premises category.

DISCUSSION

This study has examined the performance of a range of machine learning algorithms to predict the quality of drinking water sources in Ethiopia from household survey data.

The models performed well in predicting *E. coli* contamination at point of collection. RF performed the best across most metrics with XGBoost a close runner up. Overall, predictions for ESS3

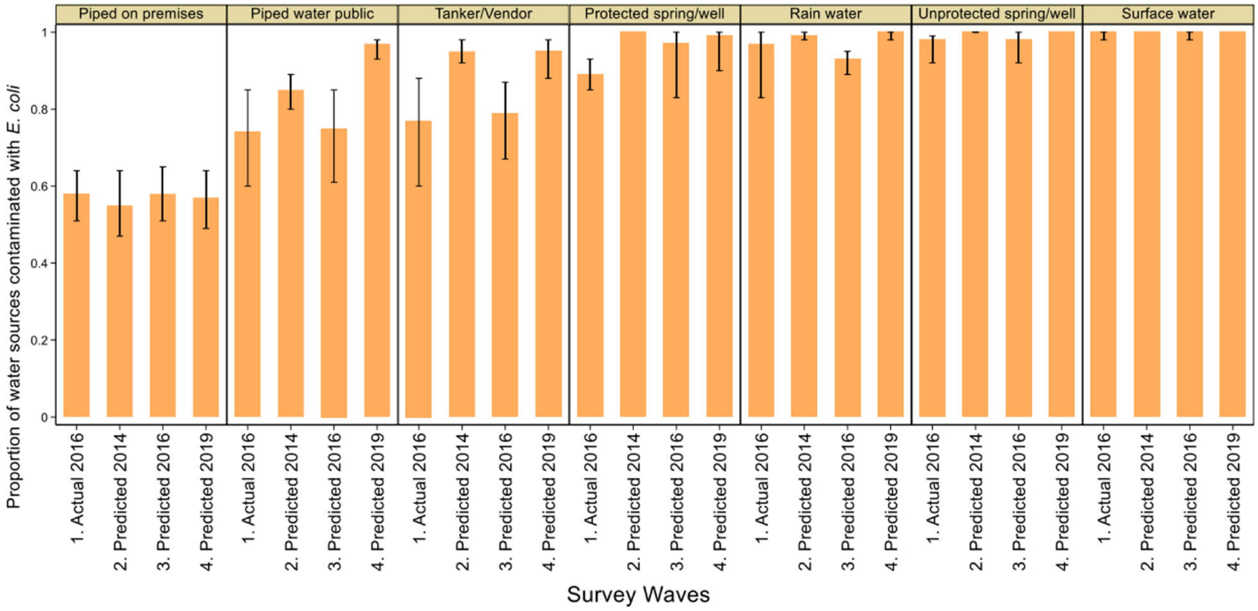

Fig. 4 Prediction results by water source type. Prediction results of *E. coli* contamination of drinking water sources in three survey waves by water source type. Error bars represent 95% confidence intervals.

(2015/16 ESS) were comparable to the actual data and the proportion of misclassified results was low (2.4%). There was some variation in model performance by water supply type and by residence. Among water source types, misclassification rates were highest for tanker truck/vendor and piped water. Water quality in piped supplies and truck/vendor water is highly dependent on the management which may be poorly predicted by the available variables. Notably, we were unable to include operational parameters such as chlorine residual in our models as this information was only available for a subset ($n = 1297$) of households in wave 3. In contrast, misclassification was very uncommon for protected wells and springs, unprotected wells and springs, rainwater, and surface water, reflecting the very low chance of these water sources not containing *E. coli* in Ethiopia. The misclassification rate was lower in rural (1.6%) than urban (5.1%) areas, again reflecting the higher risk of contamination in rural areas.

Examining variable importance from the RF model we find that the top five predictors were: distance to nearest market, mean rainfall, mean temperature, source type—piped own, and open defecation in the EA. That water source type is among the top predictors is expected given the considerable variations in quality by water source type observed in the ESS water quality module and in other countries[2]. For example, a systematic review of water quality studies in LMICs found piped water to be considerably less likely to be contaminated than other types of improved drinking water[24]. Notably, many of the top features are either basic geospatial information generated using the household's location or socioeconomic variables that are routinely captured in many household surveys such as DHS, LSMS and MICS.

We examined the relative performance of RF models across scenarios depending on the types of data available. This analysis was conducted to understand the contributions water source type, household-level and geospatial variables to model performance. In comparison with the reference "full features model", the scenario relying solely on information about the type of water source performed poorly (AUC = 0.80 vs 0.91). This is expected given the wide range of reported contamination rates for users of the same water source type as observed in Ethiopia and elsewhere[2,25]. Scenarios that relied only on geospatial information (AUC = 0.91) or household-level variables (AUC = 0.88) performed surprisingly well with performance equivalent to or approaching that of the reference scenario. This finding suggests that predictive models relying only on geospatial variables may suffice for the purposes of understanding variations in risk of *E. coli* contamination at the point of collection and support the use of machine learning to generate risk maps[13]. Similarly, the finding that household-level variables alone can provide good predictive performance. The finding suggests that the approach taken in this study could be applied to the 25 + MICS that are not disseminated with GPS coordinates[2].

Lastly, we predicted water quality at the point of collection for households in wave 2 (the 2013/14 ESS) and wave 4 (the 2018/19 ESS). The results provided a similar picture to that of the reference data. The findings of this exercise, however, cannot be assessed against any "truth" which is a key limitation here. Further work is needed to examine the performance of these predictive models over time which may necessitate the incorporate variables that capture temporal changes, including seasonality[26]. At present there are few countries that have repeated water quality testing in household surveys. Bangladesh and Nepal MICS would be good candidates for this analysis, especially given we have found good performance in the scenarios excluding geospatial variables which are not currently available in MICS.

A key strength of this analysis is the use of common machine learning techniques and a workflow in R that can be adapted for use in other household surveys. Here we utilized Boruta to automate the selection of features to include in the model. Future studies could examine the potential benefits of alternative approaches, including the use of ensemble methods. There are a number of limitations to the study and the underlying dataset. Firstly, *E. coli* measurements are a "snapshot" and consequently reflect the levels of contamination at the time samples were collected from households. As a measure of water quality, *E. coli* is not a perfect proxy for fecal contamination – for example being more sensitive to chlorine than pathogens such as cryptosporidium[27]. Secondly, the ESS water quality module was administered by separate teams revisiting households from the wave 3 of the survey. There is some discordance between the types of water sources reported in wave 3 and those from the water quality survey. For example, there is about a 10 percentage points difference in access to water from improved sources[28]. The differences in reporting on types of water source used between these two assessments may partially explain the greater contamination predicted in waves 2 and 4 and could be the result of seasonal or multiple source use[29]. Third, we did not include all variables collected in the ESS and there is inherently a choice in the geospatial datasets to consider in the analysis. In addition, emphasis is given to variables that are commonly available in different household surveys. The selection of which features to include and the pre-processing decisions introduce a degree of subjectivity and may influence the resulting performance of machine learning models. Fourth, our study examined drinking water quality at the point of collection and the performance of machine learning models (and relative importance of different sets of features) may differ for *E. coli* contamination at the point of use (i.e., immediately prior to consumption).

METHODS

Input data

The analytical framework begins with input data and continues to data preparation, modeling and application (Fig. 5). The study uses the ESS. The survey is a collaboration between the Central Statistics Agency and the World Bank under the Living Standards Measurement Study- Integrated Surveys on Agriculture (LSMS-ISA) project. ESS began in 2011/12 and the first wave, ESS1 covered rural and small-town areas. The survey was expanded to include medium and large towns in 2013/14 (ESS2). The 2013/2014 sample households were again visited in 2015/16 (ESS3) during which the water quality module was implemented. The survey was fielded again in 2018/19 (ESS4) with a refreshed sample. This study is primarily based on the 2016 Survey (ESS3) and associated water quality survey[18,28]. In this study, ESS2 is the Earlier Survey, ESS3 is the Reference Survey, and ESS4 is the Latest Survey. ESS1 was not used because the survey did not cover medium and large towns. See the Data Availability section for further information on these data sources including metadata.

ESS is a multi-topic household survey with several individual and household level socioeconomic and demographic information. These included basic individual-level demographic information on household structure, education, health, and labor market outcomes, as well as several household-level information such as household assets, consumption expenditure, dwelling characteristics, access to electricity, water, and sanitation facilities. ESS data also comes with a range of geospatial variables that are constructed by mapping the household's location to other data available for the area. These include, among other things, rainfall, temperature, greenness, wetness, altitude, population density, the household's closeness to the nearest major road, urban and market centers. In addition, the 2015/16 survey (ESS3) which is the main focus of this study, implemented a water quality module that included microbial and chemical tests to measure water quality. The microbial test included the presence of *E. coli*, WHO's preferred indicator of fecal contamination[5].

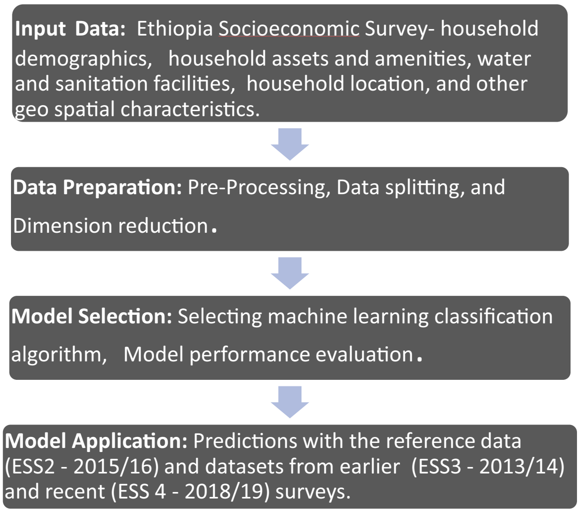
Input Data: Ethiopia Socioeconomic Survey- household demographics, household assets and amenities, water and sanitation facilities, household location, and other geo spatial characteristics.
Data Preparation: Pre-Processing, Data splitting, and Dimension reduction.
Model Selection: Selecting machine learning classification algorithm, Model performance evaluation.
Model Application: Predictions with the reference data (ESS2 - 2015/16) and datasets from earlier (ESS3 - 2013/14) and recent (ESS 4 - 2018/19) surveys.

Fig. 5 Analysis Framework. Methodological workflow from input data to model application.

Response variable

The response variable in this study is the presence of *E. coli* contamination at the point of collection. Contaminated drinking water refers to the detection of *E. coli* in water samples collected from the household's drinking water source.

Data preparation

The objective of this study was to develop a predictive model for drinking water contamination from minimal socioeconomic information. Therefore, only features that are often included in household surveys are considered. For example, the 2015/16 water quality module has some information on the chemical and physical characteristics of the water. These variables were not included in the training dataset because they are not usually available in other surveys. Therefore, the data preparation for this study considered only selected variables.

Data preparation activities included pre-processing, data splitting, and dimension reduction. The pre-processing step involved constructing some variables from existing variables, variable transformation, and treating missing values by imputation or dropping them from the analysis. Constructed variables included wealth index and open defecation in the area. The wealth index was constructed from selected assets using principal component analysis. Open defecation in the area is an enumeration area (EA) level variable and indicates the proportion of households in the EA who do not have a toilet facility. Variables that were transformed include the water source type. For example, we combined boreholes, protected springs and wells into a single category given the comparatively low number of respondents and in order to harmonize responses across the three waves of the survey. Similarly, unprotected springs and wells were combined. Consequently, the water source type list included in the model selection analysis had fewer categories than in the raw data.

To assess how the classifiers generalize to unseen data, the pre-processed data was split into training and test datasets stratified by the distribution of the response variable. Accordingly, 80% of the data is assigned to the training dataset and the remaining 20% is assigned to the test dataset. The training dataset was used to train the classifiers and estimate the hyperparameters, and the test dataset was used to evaluate the performance of the classifiers and get an independent assessment of how well the classifiers performed in predicting the positive class (contaminated drinking water source). To reduce the dimension of the processed data, the Boruta feature selection algorithm was used. The final list of features used in the analysis is presented in Supplementary Table 1.

Statistical analysis

We examined a few commonly used classification algorithms including GLM, GLMNET, KNN, SVM, and two decision tree-based classifiers: RF, and XGBoost. To obtain the optimal values of the classifiers' hyperparameters that maximize the area under the ROC, we tuned the non-liner classifiers using regular grid search method.

The GLM uses a parametric model allowing for different link functions for the response variable. For classification purposes, the response values are categorical. Especially in this study, we have a binary classification problem; i.e., "contaminated" versus "non-contaminated". Therefore, logistic regression is used as a reference model. The glm R package was used in this study[30].

The GLMNET classifier uses GLM via penalized maximum likelihood. The lasso and elastic net are popular types of penalized linear regression (or regularized linear regression models) that add penalties to the loss function during training. It promotes simpler models with better accuracy and removes features that are highly correlated. We also used glmnet R package for the GLMNET classifier and tuned two hyperparameters penalty (regularization parameter) and mixture (representing relative amount of penalties).

KNN is one of the most widely used non-parametric classifiers. It defines similarity as being in close proximity. In other words, it classifies a new case or data point based on its distance or closeness to the majority of its k nearest neighbor points in the training set. We used "kknn" package in R and tuned two hyperparameters *neighbors* (nearest neighbors) and *weight_func* (distance weighting function).

SVM is another classification method that uses distance to the nearest training data points. It classifies data points by using hyperplanes with the maximum margin between classes in high dimensional feature space[31]. It works for cases not linearly separable. In this study, we used a non-linear kernel ("kernlab") package in R and tuned two hyperparameters including *cost* and *degree* (polynomial degree).

RF is an ensemble method that builds multiple decision trees by sampling the original data set multiple times with replacement[32]. Therefore, it uses a subset of the original dataset to train the decision trees and to separate different classes as much as possible. RF combines the trees at the end by taking the majority of votes from those trees. Although large number of trees will slow the process, the greater number of trees in the forest help improve the overall accuracy and prevent the problem of overfitting. We used "ranger" package in R, which provides the importance of features as well. We tuned the following three hyperparameters: *mtry* (number of randomly selected predictors), *min_n* (minimal node size), and *trees* (1000).

XGBoost is another machine learning ensemble method which uses the gradient of a loss function that measures the performance[33]. Different than other ensemble methods, which train models in isolation of one another, XGBoost (or boosting) trains models sequentially by training each new model to correct the errors made by the previous ones. This continues until there is no scope of further improvements. XGBoost is fast to execute in general and gives good accuracy. In this study, we used "XGBClassifier" from "xgboost" package in R. The xgboost package has few tunable parameters and we tuned two of them: *trees* (trees) and *tree_depth* (tree depth).

The classification algorithms are evaluated using metrics that are calculated from the four predicted results of the confusion matrix: (i) true positive (TP) or correctly predicted as contaminated, (ii) true negative (TN) or correctly predicted as not contaminated, (iii) false positive (FP) or wrongly predicted as contaminated, and (iv) false negative (FN) or wrongly predicted as not contaminated. With our data being class-imbalanced, we used a combination of metrics to evaluate the models. We calculated accuracy, sensitivity (also known as recall or true positive rate (TPR)), specificity or true negative rate (TNR), F1 score, and area under the curve (AUC) of Receiver Operating Characteristics (ROC). The positive cases are more important than the negative cases and the goal is to make sure the best performing model maximizes the TPR. Finally, given the data we used is of imbalanced classes we have implemented resampling techniques[17]. These include upsampling the minority class and downsampling the majority class (See Supplementary Tables 3 and 4). However, there were no significant improvements in the prediction results. The AUC for the RF model using upsampling and downsampling techniques is 0.90 (95% CI 0.88, 0.93). Similarly, AUC for the XGBoost model is 0.90 (95% CI 0.87, 0.92) for upsampling and 0.89 (95% CI 0.86, 0.92). These are similar to the main results reported in Table 2.

The analyses were conducted with the R programming language.

DATA AVAILABILITY

The raw data that support the findings of this study are available in World Bank Microdata Library at the following sites.
ESS Wave 2: https://microdata.worldbank.org/index.php/catalog/2247/get-microdata
ESS Wave 3: https://microdata.worldbank.org/index.php/catalog/2783/get-microdata
ESS Wave 4: https://microdata.worldbank.org/index.php/catalog/3823/get-microdata

CODE AVAILABILITY

The data analysis code used in this study is available from the authors upon request.

Received: 19 January 2022; Accepted: 1 August 2023;
Published online: 08 September 2023

REFERENCES

1. WHO/UNICEF. Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs. Geneva: World Health Organization (WHO) and the United Nations Children's Fund (UNICEF) (2021).
2. Bain, R., Johnston, R., Khan, S., Hancioglu, A. & Slaymaker, T. Monitoring drinking water quality in nationally representative household surveys in low- and middle-income countries: cross-sectional analysis of 27 multiple indicator cluster surveys 2014–2020. *Environ. Health Perspect.* **129**, 097010 (2021).
3. Prüss-Ustün, A. et al. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: an updated analysis with a focus on low- and middle-income countries. *Int. J. Hyg. Environ. Health* **222**, 765–777 (2019).
4. United Nations. SDG Indicators - Metadata repository, https://unstats.un.org/sdgs/metadata/files/Metadata-06-01-01.pdf (2017).
5. WHO. Guidelines for drinking-water quality. (2017).
6. Jean, N. et al. Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
7. Embaye, W. T., Zereyesus, Y. A. & Chen, B. Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: evaluations of hedonic pricing and machine learning approaches. *PloS One* **16**, e0244953 (2021).
8. Razzaq, A. et al. An automatic determining food security status: machine learning based analysis of household survey data. *Int. J. Food Prop.* **24**, 726–736 (2021).
9. Wang, S., Azzari, G. & Lobell, D. B. Crop type mapping without field-level labels: random forest transfer and unsupervised clustering techniques. *Remote Sens. Environ.* **222**, 303–317 (2019).
10. Lobell, D. B. et al. Eyes in the sky, boots on the ground: assessing satellite- and ground-based approaches to crop yield measurement and analysis. *Am. J. Agric. Econ.* **102**, 202–219 (2020).
11. Bonilla Cedrez, C., Chamberlin, J., Guo, Z. & Hijmans, R. J. Spatial variation in fertilizer prices in Sub-Saharan Africa. *PloS One* **15**, e0227764 (2020).
12. Jeandron, A., Cumming, O., Kapepula, L. & Cousens, S. Predicting quality and quantity of water used by urban households based on tap water service. *npj Clean Water* **2**, 23 (2019).
13. Poulin, C. et al. What environmental factors influence the concentration of fecal indicator bacteria in groundwater? Insights from explanatory modeling in Uganda and Bangladesh. *Environ. Sci. Technol.* **54**, 13566–13578 (2020).
14. Podgorski, J. & Berg, M. Global threat of arsenic in groundwater. *Science* **368**, 845–850 (2020).
15. Podgorski, J. E., Labhasetwar, P., Saha, D. & Berg, M. Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environ. Sci. Technol.* **52**, 9889–9898 (2018).
16. Chen, K. et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **171**, 115454 (2020).
17. Bourel, M. et al. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Res.* **202**, 117450 (2021).
18. Central Statistics Agency & World Bank. Ethiopia Socioeconomic Survey 2015-2016. https://doi.org/10.48529/ampf-7988 (2017).
19. Central Statistics Agency & ICF. Ethiopia Demographic and Health Survey 2016. Ethiopia Demographic and Health Survey 2016 [FR328] (dhsprogram.com) (2017).
20. Bain, R. et al. Fecal contamination of drinking-water in low- and middle-income countries: a systematic review and meta-analysis. *PLoS Med.* **11**, e1001644 (2014).

21. Kumpel, E., Peletz, R., Bonham, M. & Khush, R. Assessing drinking water quality and water safety management in Sub-Saharan Africa using regulated monitoring data. *Environ. Sci. Technol.* **50**, 10869–10876 (2016).
22. WHO/UNICEF. Core questions on water, sanitation and hygiene for household surveys: 2018 Update. (2018).
23. Welle, K., Schaefer, F., Butterworth, J. & Bostoen, K. Enabling or disabling? Reflections on the Ethiopian National WASH inventory process. *IDS Bull.* **43**, 44–50 (2012).
24. Shields, K. F., Bain, R. E., Cronk, R., Wright, J. A. & Bartram, J. Association of supply type with fecal contamination of source water and household stored drinking water in developing countries: a bivariate meta-analysis. *Environ. Health Perspect.* **123**, 1222–1231 (2015).
25. Yang, H. et al. Water safety and inequality in access to drinking-water between rich and poor households. *Environ. Sci. Technol.* **47**, 1222–1230 (2013).
26. Kostyla, C., Bain, R., Cronk, R. & Bartram, J. Seasonal variation of fecal contamination in drinking water sources in developing countries: a systematic review. *Sci. Total Environ.* **514**, 333–343 (2015).
27. Charles, K. J., Nowicki, S. & Bartram, J. K. A framework for monitoring the safety of water services: from measurements to security. *npj Clean Water* **3**, 36 (2020).
28. Central Statistics Agency. Drinking Water Quality in Ethiopia: Results from the 2016 Ethiopia Socioeconomic Survey. (Addis Ababa, 2017).
29. Daly, S. W., Lowe, J., Hornsby, G. M. & Harris, A. R. Multiple water source use in low- and middle-income countries: a systematic review. *J. Water Health* **19**, 370–392 (2021).
30. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
31. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
32. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
33. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System, In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794 (2016).

AUTHOR CONTRIBUTIONS

A.A., R.B., T.D., A.D., R.J. and T.S. contributed to conceptualization. T.D. prepared the data. A.A., R.B., A.D. and T.D. outlined the methodology. A.A. and T.D. conducted the formal analysis. A.A., R.B. and A.D. wrote the original draft. A.A., R.J. and T.S. provided supervision. A.A., R.B., T.D., A.D., R.J. and T.S. reviewed and edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41545-023-00272-8.

Correspondence and requests for materials should be addressed to Alemayehu A. Ambel.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.