




Article

Application of Machine Learning to Identify Influential Factors for Fecal Contamination of Shallow Groundwater

Jianyong Wu ^{1,*}, Yanni Cao ¹, Md. Sirajul Islam ² and Michael Emch ^{3,4}

¹ Division of Environmental Health Sciences, College of Public Health, The Ohio State University, Columbus, OH 43210, USA; cao.1637@osu.edu

² International Centre for Diarrhoeal Disease Research, Dhaka 1212, Bangladesh; sislam@icddr.org

³ Department of Geography and Environment, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; emch@unc.edu

⁴ Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

* Correspondence: wu.6255@osu.edu; Tel.: +1-614-292-3435

Abstract: Understanding influential factors for fecal contamination in groundwater is critical for ensuring water safety and public health. The objective of this study is to identify key factors for fecal contamination of shallow tubewells using machine learning methods. Three methods, including recursive feature elimination (RFE) with XGBoost, Random Forest, and mutual information, were implemented to examine *E. coli* presence and concentration in 1495 tubewell water samples in Matlab, Bangladesh. For *E. coli* presence, climatic variables, including average rainfall and temperature over the 30, 15, and 7 days preceding sampling, as well as ambient temperature and rainfall on the sampling day, emerged as critical predictors. Land cover characteristics, such as the percentages of urban and agricultural areas within 100 m of a tubewell, were also significant. For *E. coli* concentration, land cover characteristics within 100 m, the number of hot and heavy-rain days in the 30 days preceding sampling, average rainfall and temperature in the 3 days preceding sampling, and ambient temperature on the sampling day were identified as key drivers. Random Forest and mutual information yielded results that were more similar to each other than to those of RFE with XGBoost. The findings highlight the interplay between climatic factors, land use, and population density in determining fecal contamination in shallow well water and demonstrate the power of machine learning algorithms in ranking these factors.

Keywords: fecal contamination; *E. coli*; climate extreme; groundwater; global health; Bangladesh; XGBoost; random forest; mutual information; feature selection



Academic Editor: Constantinos V. Chrysikopoulos

Received: 7 December 2024

Revised: 3 January 2025

Accepted: 6 January 2025

Published: 9 January 2025

Citation: Wu, J.; Cao, Y.; Islam, M.S.; Emch, M. Application of Machine Learning to Identify Influential Factors for Fecal Contamination of Shallow Groundwater. *Water* **2025**, *17*, 160. <https://doi.org/10.3390/w17020160>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Groundwater is a major source of drinking water in many areas of the world [1–3]. It is estimated that approximately 2.2 billion people in the world use groundwater for daily consumption [4]. Groundwater is assumed to have better microbial water quality than surface water due to soil and bedrock barriers [5], often receiving little or no treatment preceding drinking [6]. However, groundwater quality is susceptible to contamination from human activities, natural geochemical processes, and land use changes, which can compromise its suitability for drinking and other purposes [7,8]. Effective extraction and utilization necessitate regular monitoring of water quality, the implementation of treatment technologies to remove contaminants, and management practices to prevent over-extraction and pollution [9]. Furthermore, modeling and predicting groundwater quality is essential

to ensure sustainable use and minimize the risk of contaminant exposure, particularly in rural areas where routine monitoring may not be feasible.

In recent decades, studies have shown that groundwater is also vulnerable to microbial contamination [10–14]. Various pathogens, including pathogenic *E. coli*, *Shigella* spp., *Vibrio cholerae*, *Campylobacter* spp., Hepatitis A, norovirus, rotavirus, *Giardia lamblia*, and *Cryptosporidium*, have been detected in groundwater and associated with waterborne disease outbreaks [4,15–17]. To address this concern, it is important to understand the factors that influence the level of fecal contamination in groundwater. First, identifying key factors can enable targeted risk assessments in the areas where tubewells are a primary drinking water source. Second, identifying key factors helps develop evidence-based strategies to protect groundwater quality and mitigate contamination risks. Furthermore, identifying key factors provides a basis for building models to predict fecal contamination under various scenarios.

Several studies have attempted to identify predictors or risk factors for fecal contamination in groundwater [1,18–23]. For example, a study showed that a tubewell sanitary inspection score did not predict the microbiological water quality of tubewells in three flood-prone areas in Bangladesh [19]. In rural Bangladesh, tubewells without annular seals had a high risk of fecal contamination in their water [22]. Additionally, a significant association was observed between the frequency of *E. coli* detection in shallow tubewells and the presence of a latrine within 10 m of the well during dry seasons, but not during wet seasons [18]. In Kampala, Uganda, rapid recharge of protected shallow groundwater springs after rainfall was a major risk factor that led to groundwater microbial contamination [24]. Furthermore, land use and climate extremes could significantly contribute to fecal contamination of shallow tubewells [20]. These studies examined the influence of tubewell characteristics, climatic factors, and neighborhood environments, such as land use and proximity to nearby latrines, on fecal contamination in groundwater. However, no study has evaluated the relative importance of these factors when considered collectively.

Recently, machine learning has been successfully applied in the prediction of microbial contamination in groundwater [23,25,26]. Machine learning is a branch of artificial intelligence that focuses on developing algorithms and models that allow computers to learn patterns and make predictions or decisions based on data. Machine learning algorithms can be generally categorized into supervised learning and unsupervised learning based on whether the data are labeled [27]. In supervised learning, labeled data are provided, allowing the algorithm to learn the relationship between inputs and outputs. These algorithms are often used for classification and regression [27]. Common algorithms in this category include support vector machines (SVMs), k-nearest neighbors (KNNs), decision tree (DT), Random Forest (RF), artificial neural networks (ANNs), eXtreme Gradient Boosting (XGBoost), and Naive Bayes. In contrast, unsupervised learning deals with unlabeled data, which aims to identify hidden patterns or groupings within the data. A widely used algorithm in unsupervised learning is K-means clustering, which partitions data into distinct clusters based on similarity [27]. Besides its capacity to make predictions, machine learning is also a powerful tool for selecting features, which aims to identify the most relevant variables from a dataset. By reducing dimensionality, feature selection helps to focus on the most influential factors while minimizing noise and computational cost. Techniques such as recursive feature elimination (RFE), the Random Forest importance index, and mutual information score are commonly used for this purpose, as they rank features based on their contribution to the predictive power of the model.

To date, studies that used these techniques to rank important factors for fecal contamination in groundwater are still unavailable. Therefore, the objective of this study is to apply machine learning algorithms to identify the key factors influencing fecal contamination

in shallow groundwater. First, a wide range of potential factors affecting fecal contamination in shallow tubewells were collected. Next, three machine learning approaches were employed to rank the factors influencing both the presence and concentration of *E. coli* (Figure 1). By identifying and ranking the most critical factors contributing to fecal contamination in groundwater, this study provides novel insights into understanding the complex interactions and relative importance of these factors using machine learning, thereby offering a data-driven foundation for prioritizing effective mitigation strategies.

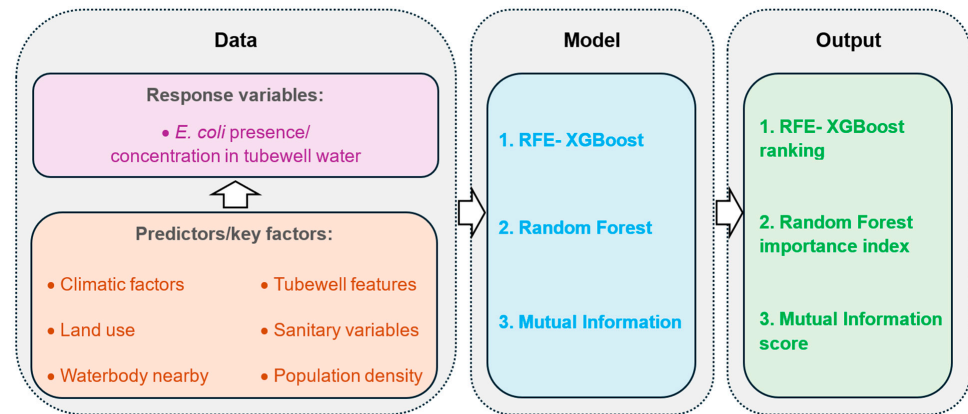


Figure 1. Flowchart illustrating the study design and methodology employed in this research.

2. Materials and Methods

2.1. Study Sites and Sampling

The study area is located in Matlab, a rural area in Bangladesh about 57 km southeast of the capital city, Dhaka (Figure 2). Matlab is one of the field sites of the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), which maintains rich datasets related to health and demographic information through its Health and Demographic Surveillance System [28]. Almost all of the people living in Matlab use groundwater as their drinking water source. We selected six villages (Barahaldia, Sardarkandi, Shakhari para, Farazikandi, Namapara, and Shankibhanga) in Matlab to measure the microbial water quality of groundwater [29]. Water samples from 92 tubewells in these villages were collected quasi-monthly from May 2008 to October 2009 to measure *E. coli*, as well as other water quality variables, including total coliforms, pH, and dissolved oxygen (DO).

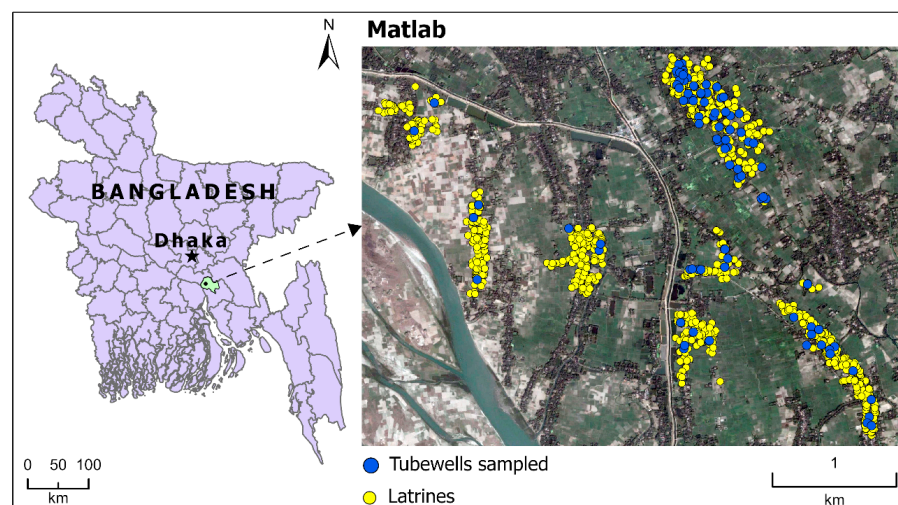


Figure 2. The study area and sampling sites in Bangladesh. The satellite image is from Google Earth [20].

2.2. Measurement of Microbial Contaminants

Groundwater samples from the 92 tubewells were tested for *E. coli* and total coliforms within 8 h of sample collection using the IDEXX Colilert-18 method [29]. Briefly, a 100 mL water sample is mixed with Colilert-18 reagent and then poured into a Quanti-Tray 2000 (IDEXX Laboratories, Inc., Westbrook, ME, USA). A separate tray is used for each groundwater sample. The tray is sealed and incubated overnight at 35 °C. Trays are placed under a UV light at 360 nm, and the number of wells that appear positive for each bacterial indicator is used to calculate the most probable number (MPN) of *E. coli* and total coliforms per 100 mL water sample. The concentrations of *E. coli* and total coliforms were measured in duplicate samples. Water temperature, pH, specific conductance, Oxidation–Reduction Potential (ORP) and DO were measured with YSI sensors (YSI, Yellow Springs, OH, USA) during sample collection.

2.3. Survey of Tubewells and Sanitary Facilities

Water and sanitation infrastructure around monitored tubewells was surveyed in 2008 and 2009. Geographical locations were measured with Trimble GeoXH global positioning system (GPS) receivers (Trimble Inc., Westminster, CO, USA) and mapped with submeter accuracy. Locations of households, latrines, and ponds within 200 m of these wells were recorded. Details about tubewell depth, installation year, and drinking water use were provided by residents, while pond characteristics (depth, use, and effluent reception) were documented and categorized into bathing, fish-farming, latrine, or no-purpose ponds [30]. The status of the tubewell platform was also recorded and wells were classified as having an intact concrete platform vs. no platform or broken platform. Latrines were classified as sanitary if there was no visible leaking effluent and the septic tank was intact. Otherwise, it was classified as an unsanitary latrine. Waterbodies near a tubewell were categorized as river, pond, ditch, and other.

2.4. Land Use and Weather Data

Land use data were generated using an object-based classification method to classify a satellite image with a resolution of 2.1 m in the study area. Detailed information about satellite data and image processing was described previously [20]. Land use was classified into six major types: agriculture, barren land, developed (urban) land, tree cover, water, and wetland. The percentage of land use types within a 100 m buffer surrounding each well was estimated. The 100 m buffer was chosen based on previous research findings that the land type within 100 m of a tubewell significantly influences fecal contamination of tubewell water [20]. Daily precipitation data were obtained from the TRMM (Tropical Rainfall Measuring Mission) online visualization and Analysis System (TOVAS). Daily average temperature data were obtained from the National Climate Data Center (NCDC). The threshold for extreme weather events (hot days and heavy-rainfall days) was defined using the 90th-percentile values during a 10-year period (from 1 January 2001 to 31 December 2010) [31]. If the daily temperature or precipitation is above the threshold, that day is defined as a hot day or a heavy-rain day. The average precipitation and the number of heavy-rain days in the 3, 7, 15, or 30 days preceding a sampling date were also calculated because these variables might be related to the level of fecal contamination in tubewell water [20].

2.5. Data Processing

After data were collected, they were compiled together using a geographic information system (GIS). The distances between tubewells and nearby latrines and between tubewells and nearby waterbodies were calculated using ArcGIS 10.1 (ERSI, Redlands, CA, USA).

Further, the total number of sanitary and unsanitary latrines within a 100 m buffer surrounding a tubewell was calculated. Population density within a 100 m buffer of a tubewell was estimated as the number of people living within 100 m of the tubewell divided by the area of the buffer.

Two dependent variables (also called response variables) were used in this study, including the presence of *E. coli* and the concentration of *E. coli*. For each variable, there are 1495 observations (data points). To identify key predictors for fecal contaminants, we obtained data for over 60 variables in 6 major categories, including water quality variables (the concentrations of total coliforms and DO, and water pH and water temperature), land use variables (the percentages of six types of land use classes), weather variables (daily precipitation and temperature), tubewell characteristics (the age, depth, and platform type of a tubewell), sanitation variables (the number of unsanitary latrines and total latrines within 100 m), and population density. Water quality variables were excluded as predictors for fecal contamination because they are not always easily obtained. Pearson correlation analysis was conducted to examine the relationship between predictors and microbial contaminants.

2.6. Identifying Influential Factors Using Recursive Feature Elimination with XGBoost

RFE with the XGBoost algorithm was employed to identify influential factors contributing to *E. coli* concentrations and presence in tubewell water. RFE is a feature selection method that iteratively removes the least essential predictors based on model performance, ranking variables according to their relative importance [32]. XGBoost was chosen for its efficiency and ability to capture complex relationships between variables [33]. For *E. coli* concentrations, we log-transformed the concentrations and used this as the response variable in the RFE-XGBoost process and ranked these 58 predictive variables. The same RFE-XGBoost approach was applied to ranking important factors for *E. coli* presence. The model was implemented with the 'xgboost' package (Version 2.1.2) in the Python 3.11 environment. 'XGBClassifier' and 'XGBRegressor' were used to rank key factors of *E. coli* presence and concentration, respectively.

2.7. Identifying Influential Factors Using the Importance Index from Random Forest

The RF algorithm was employed to identify influential factors contributing to *E. coli* concentrations and presence in tubewell water. RF is a tree-based machine learning method which makes decisions based on multiple decision trees [34]. Each tree is trained on a random subset of the data and predictor variables. RF has strengths in handling high-dimensional and non-linear data [35,36]. The importance index from RF quantifies the contribution of each predictor to the model's predictive performance based on metrics such as mean decrease in impurity or mean decrease in accuracy [37]. For *E. coli* concentrations, the log-transformed concentrations were used as the response variable and fit an RF model with the 58 predictors. The variables were then ranked based on their importance scores. Similarly, for *E. coli* presence, a binary response variable, we fit another RF model with the same 58 predictors and ranked them by their importance scores. The RF model was implemented with the 'Sklearn' package (Version 1.3.0) in the Python 3.11 environment. 'RandomForestClassifier' and 'RandomForestRegressor' were used to rank key factors of *E. coli* presence and concentration, respectively.

2.8. Identifying Influential Factors Using Mutual Information

To identify the influential factors contributing to fecal contamination in shallow tubewells, the mutual information (MI) approach, a non-parametric method, was applied to quantify the dependency between two variables [38,39]. MI captures both linear and non-

linear relationships, making it particularly effective in complex systems. The MI score is calculated using the below function [40]:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$

where X and Y are two variables, $p(x, y)$ represents the joint probability distribution of variables X and Y , and $p(x)$ and $p(y)$ are their marginal probability distributions.

To identify influential factors for *E. coli* concentration, log-transformed concentrations were used as the response variable. The MI scores were calculated for 58 potential predictor variables, and these were ranked based on their scores. Similarly, to determine influential factors for *E. coli* presence in tubewell water, a binary response variable was used to indicate the presence or absence of *E. coli*, and the MI scores were again calculated and ranked for the same 58 variables. The MI was implemented with the ‘sklearn’ package (Version 1.3.0) in the Python 3.11 environment. ‘mutual_info_classif’ and ‘mutual_info_regression’ were used to calculate MI scores of key factors for *E. coli* presence and concentration, respectively.

3. Results

3.1. Exploratory Data Analysis

A total of 1495 *E. coli* measurements were performed for samples from 92 tubewells. *E. coli* was detected in 646 measurements, with the largest concentration of 30,000 MPN/L. Descriptive statistics for these variables (e.g., mean, standard deviation, and data type) are shown in Table 1. The relationship between *E. coli* and these variables was preliminarily examined using Pearson correlation analysis (Table 2). The results show that *E. coli* presence had a significant positive correlation ($R > 0, p < 0.05$) with the number of latrines nearby (both sanitary and unsanitary), population density, the percentages of urban land and water areas, heavy-rain days, the number of heavy-rain days in 3 or 30 days preceding sampling, and average temperature and rainfall in 7 days preceding sampling. However, it was negatively correlated with the type of platform and the percentage of agricultural area ($R < 0, p < 0.05$). Similarly, *E. coli* concentration was significantly and positively correlated with the number of latrines nearby, population density, the percentages of urban land and water areas, heavy-rain days, the number of heavy-rain days in 30 days preceding sampling, and average rainfall in 7 days preceding sampling ($R > 0, p < 0.05$), but negatively correlated with the percentage of agricultural area ($R < 0, p < 0.05$). The correlation heatmap is shown in Figure 3.

Table 1. Descriptive statistics of fecal indicators and potential predictors of fecal contaminants in groundwater.

Variables (Unit) N = 1495	Mean	Standard Deviation	Min	Max	Data Types
<i>E. coli</i> presence	0.43	0.50	0	1	Categorical
<i>E. coli</i> concentration (MPN/100 mL)	38.09	248.29	0.30	3000	Continuous
The number of unsanitary latrines in 100 m	10.71	7.67	1	26	Discrete
The number of sanitary latrines in 100 m	14.87	7.24	2	32	Discrete
The type of nearby waterbody	1.59	0.68	146	1	Categorical
The type of the nearest latrine (sanitary vs. unsanitary)	0.38	0.49	0	1	Categorical
Distance to the nearest latrine (m)	13.02	8.55	0.7	40.43	Continuous
The type of platform	0.62	0.49	0	1	Categorical

Table 1. Cont.

Variables (Unit) N = 1495	Mean	Standard Deviation	Min	Max	Data Types
Well depth (m)	15.38	5.44	7.5	36	Continuous
Population density in 100 m	8.25	0.70	6.46	9.15	Continuous
The percentage of urban area in 100 m (%)	11.96	6.68	2.5	31.69	Continuous
The percentage of water area in 100 m (%)	21.49	9.32	5.72	40.09	Continuous
The percentage of agricultural land in 100 m (%)	29.54	17.89	2.92	70.40	Continuous
Heavy-rain day (Yes/No)	0.21	0.41	0	1	Discrete
Daily mean temperature (°C)	27.26	2.81	17.95	31.37	Continuous
Daily rainfall (mm)	11.66	20.60	0	108.99	Continuous
Average daily temperature in 7 days preceding sampling	26.89	3.54	17.96	30.65	Continuous
Average rainfall in 7 days preceding sampling	10.78	10.34	0	49.23	Continuous
Average daily temperature in 30 days preceding sampling	26.90	3.19	19.02	30.12	Continuous
The number of heavy-rain days in 30 days preceding sampling	26.90	3.19	19.02	30.12	Discrete
The number of heavy-rain days in 3 days preceding sampling	0.59	0.83	0	3	Discrete

Table 2. Pearson correlation between potential predictors and *E. coli* presence and concentration.

Variables	<i>E. coli</i> Presence		<i>E. coli</i> Concentration *	
	r	p	r	p
The type of platform	−0.060	0.021	−0.041	0.113
Distance to the nearest latrine (m)	0.024	0.349	0.043	0.099
The type of nearby waterbody	0.035	0.174	0.046	0.073
The number of unsanitary latrines in 100 m	0.074	0.004	0.111	<0.001
The number of sanitary latrines in 100 m	0.072	0.005	0.093	<0.001
Well depth (m)	−0.024	0.361	0.013	0.606
Population density in 100 m	0.093	<0.001	0.135	<0.001
The percentage of urban area in 100 m (%)	0.122	<0.001	0.142	<0.001
The percentage of water area in 100 m (%)	0.059	0.023	0.085	0.001
The percentage of agricultural land in 100 m (%)	−0.099	<0.001	−0.128	<0.001
Average rainfall in 7 days preceding sampling	0.103	<0.001	0.091	<0.001
Average temperature in 7 days preceding sampling	0.051	0.047	0.044	0.086
Daily rainfall	0.025	0.331	−0.003	0.900
Daily average temperature (°C)	0.037	0.154	0.025	0.325
Heavy-rain day (yes/no)	0.076	0.003	0.052	0.043
The number of heavy-rain days in 3 days preceding sampling	0.070	0.007	0.042	0.104
The number of heavy-rain days in 30 days preceding sampling	0.177	<0.001	0.148	<0.001
Average temperature in 30 days preceding sampling	0.065	0.012	0.057	0.028

Note: * *E. coli* concentration values are log-transformed for analysis.

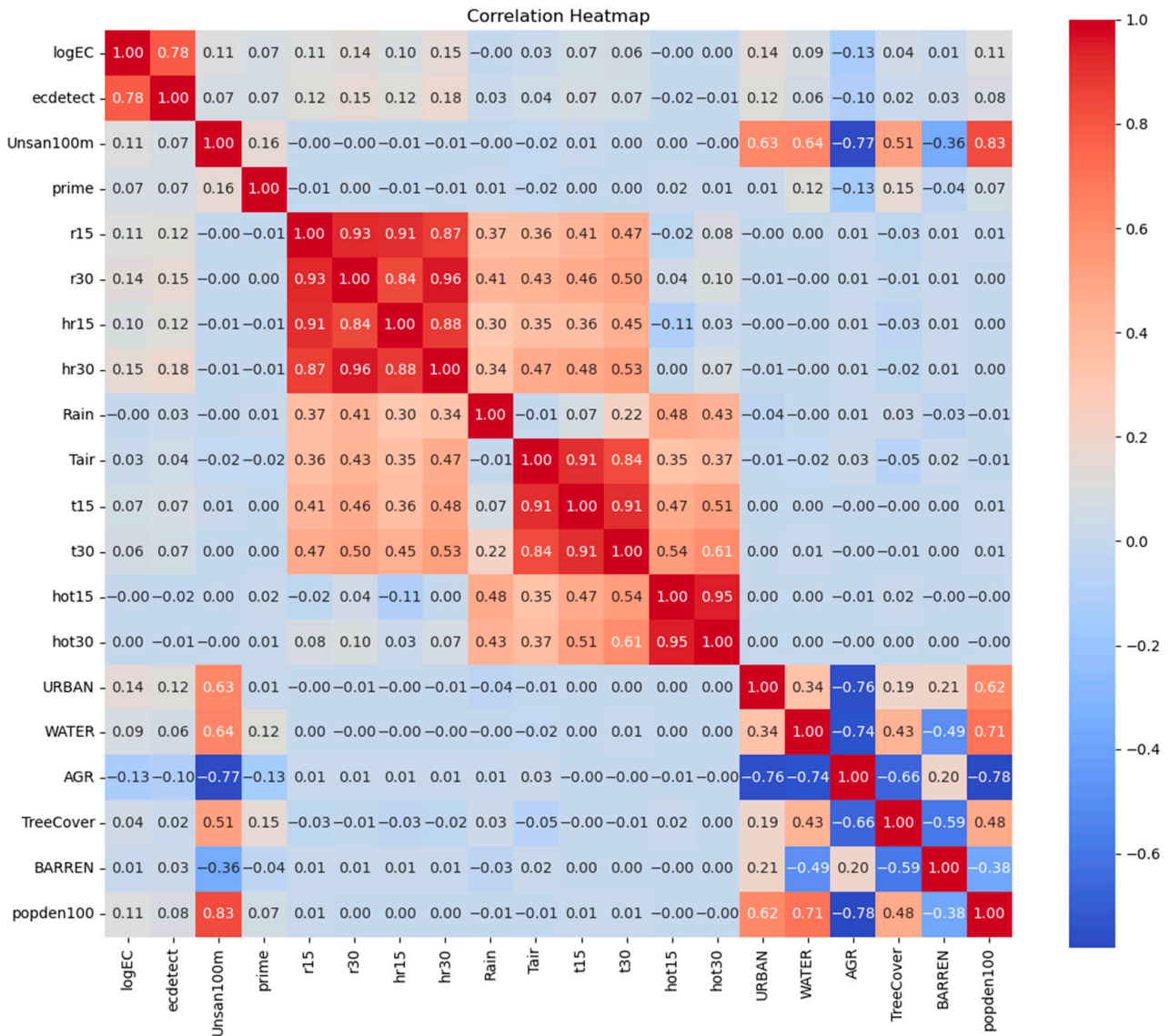


Figure 3. Correlation heatmap illustrating Pearson correlation coefficients among key variables. logEC: the log-transformed *E. coli* concentration; ecdetect: whether *E. coli* is presence in a tubewell; Unsan100m: number of unsanitary latrines within 100 m; prime: volume of water used for tubewell priming; r15: average rainfall in 15 days preceding sampling; r30: average rainfall in 30 days preceding sampling; hr15: number of heavy-rain days in 15 days preceding sampling; hr30: number of heavy-rain days in 30 days preceding sampling; Rain: rainfall on the sampling day; Tair: ambient temperature on the sampling day; t15: average temperature in 15 days preceding sampling; t30: average temperature in 30 days preceding sampling; hot15: number of hot days in 15 days preceding sampling; hot30: number of hot days in 30 days preceding sampling; URBAN: percentage of urban area within 100 m of a tubewell; WATER: percentage of water area within 100 m of a tubewell; ARG: percentage of agricultural land within 100 m of a tubewell; TreeCover: percentage of tree cover within 100 m of a tubewell; BARREN: percentage of barren land within 100 m of a tubewell; popden100: population density within 100 m of a tubewell.

3.2. Key Factors Influencing *E. coli* Presence

Using RFE with XGBoost, the factors influencing *E. coli* presence were ranked. The top-ranked factors include the average rainfall in 30 days preceding sampling, the average rainfall in 7 days preceding sampling, the volume of water used for tubewell priming, population within 200 m of a tubewell, the number of hot days in 30 days preceding sampling, the percentage of urban area within 100 m of a tubewell, the percentage of

agricultural land within 100 m of a tubewell, the average temperature in 30 days preceding sampling, population within 25 m of a tubewell, and the percentage of tree cover within 100 m of a tubewell (Table 3). According to the importance index of the RF algorithm, the top influential factors include the average rainfall in 30 days preceding sampling, the average rainfall in 15 days preceding sampling, the average temperature in 15 days preceding sampling, the ambient temperature on the date of sampling, the average temperature in 7 days preceding sampling, the average rainfall in 7 days preceding sampling, average temperature in 30 days preceding sampling, the percentage of urban area within 100 m of a tubewell, the average temperature in 3 days preceding sampling, and the average rainfall in 3 days preceding sampling (Figure 4). The MI score shows that the following factors are major factors influencing *E. coli* presence in tubewells, including the ambient temperature on the date of sampling, the average temperature in 3 days preceding sampling, the average rainfall in 30 days preceding sampling, the average temperature in 30 days preceding sampling, the percentage of agricultural land within 100 m of a tubewell, the average temperature in 15 days preceding sampling, the number of heavy-rain days in 30 days preceding sampling, the average temperature in 7 days preceding sampling, average rainfall in 7 days preceding sampling, and the number of hot days in 30 days preceding sampling (Figure 5).

Table 3. Ranking important factors using RFE with XGBoost for *E. coli* presence.

Rank	Features
1	Average rainfall in 30 days preceding sampling
2	Average rainfall in 7 days preceding sampling
3	Volume of water used for tubewell priming
4	Population within 200 m of a tubewell
5	Number of hot days in 30 days preceding sampling
6	Percentage of urban area within 100 m of a tubewell
7	Percentage of agricultural land within 100 m of a tubewell
8	Average temperature in 30 days preceding sampling
9	Population within 25 m of a tubewell
10	Percentage of tree cover within 100 m of a tubewell
11	Average temperature in 7 days preceding sampling
12	Population within 50 m of a tubewell
13	Horizontal distance from a tubewell to the nearest latrine
14	Rainfall on the sampling day
15	Number of hot days in 15 days preceding sampling
16	Ambient temperature on the sampling day
17	Number of people drinking water from a tubewell
18	Average rainfall in 15 days preceding sampling
19	Vertical distance from a tubewell to the nearest pond
20	Number of sanitary latrines within 100 m
21	Distance from a tubewell to the nearest latrine
22	Percentage of water area within 100 m of a tubewell
23	Average rainfall in 3 days preceding sampling
24	Tubewell depth
25	Types of discharge near a tubewell

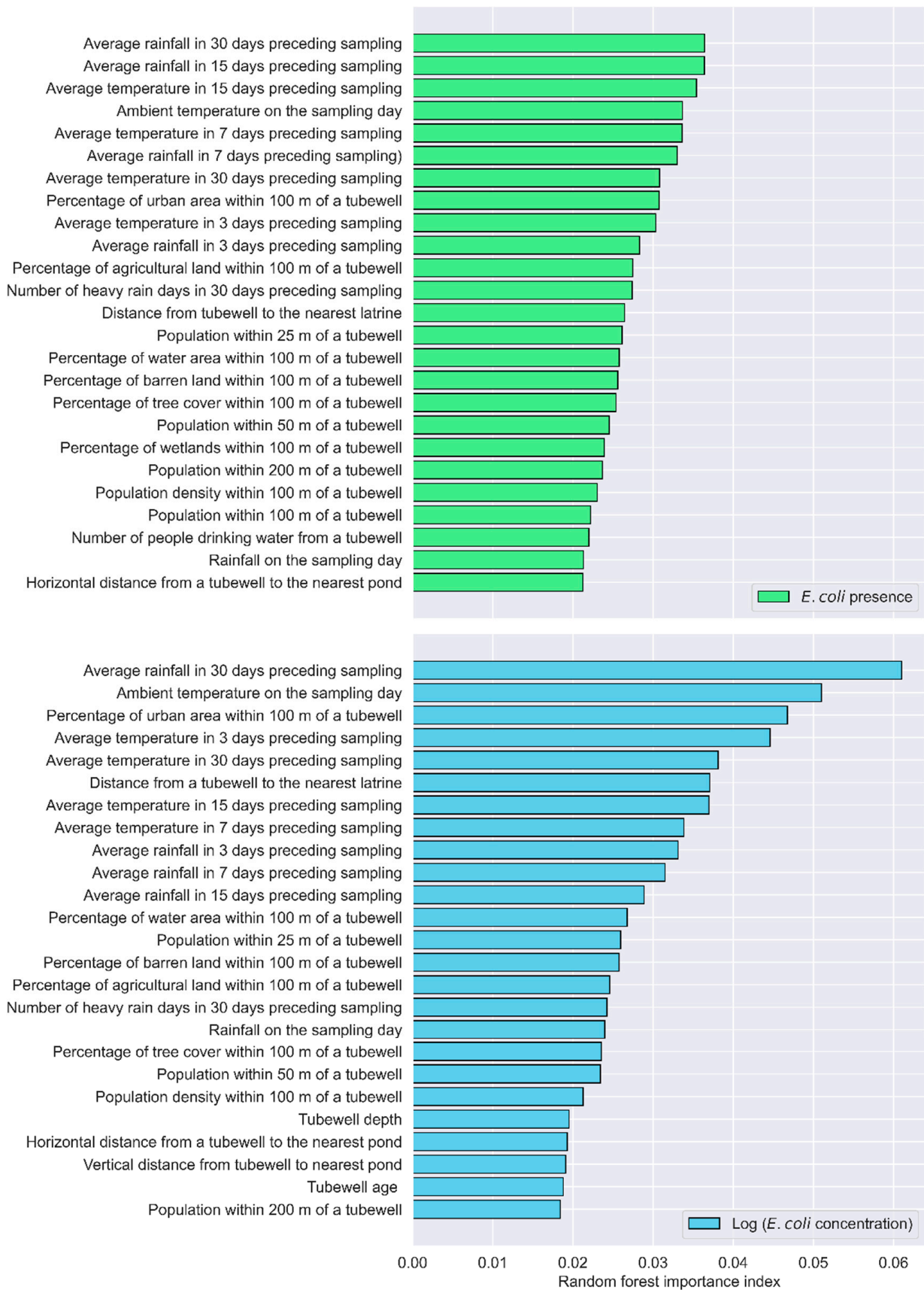


Figure 4. Key factors for *E. coli* presence and concentration based on the importance index from Random Forest.

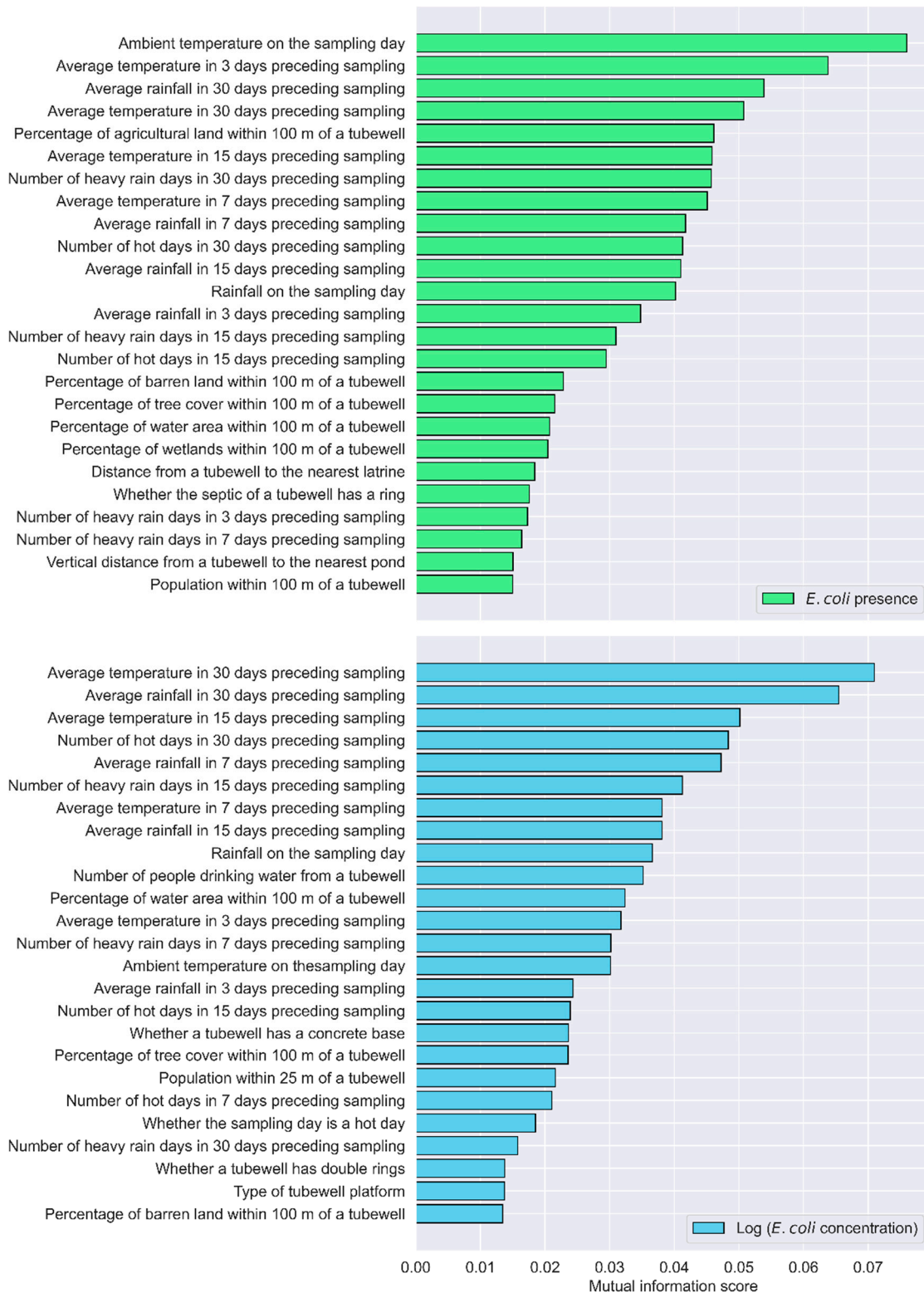


Figure 5. Key factors for *E. coli* presence and concentration based on the mutual information score.

3.3. Key Factors Influencing *E. coli* Concentration

The analysis identified a range of climatic, land use, and demographic factors as key factors influencing *E. coli* concentration in tubewells, with consistent patterns emerging across multiple feature selection methods. The results from the RFE with XGBoost approach show that the percentages of agricultural land, barren land, and urban areas within 100 m of a tubewell ranked highly. Additionally, climatic variables, such as the number of heavy-rain

days and hot days in the 30 days preceding sampling, as well as the ambient temperature on the sampling day, were important contributors. Other influential factors include the percentage of tree cover and water bodies within the same radius, along with short-term climatic metrics like average rainfall and temperature in the 3 days preceding sampling (Table 4). The RF importance index emphasized the role of cumulative climatic conditions, with average rainfall and temperature over various time frames, particularly the 30, 15, and 7 days preceding sampling, emerging as critical factors. Ambient temperature on the sampling day was also consistently important, alongside the distance from a tubewell to the nearest latrine. Land use characteristics, such as the percentage of urban area within 100 m, were again ranked among the top predictors (Figure 4). Mutual information analysis similarly underscored the importance of climatic variables, particularly the average temperature and rainfall over the 30, 15, and 7 days preceding sampling. The numbers of heavy-rain and hot days during these periods are also influential. Additionally, rainfall on the sampling day and the number of people drinking water from a tubewell are important factors identified by this method (Figure 5).

Table 4. Ranking important factors using RFE with XGBoost for *E. coli* concentration.

Rank	Features
1	Percentage of agricultural land within 100 m of a tubewell
2	Percentage of barren land within 100 m of a tubewell
3	Number of heavy-rain days in 30 days preceding sampling
4	Number of hot days in 30 days preceding sampling
5	Ambient temperature on the sampling day
6	Percentage of water area within 100 m of a tubewell
7	Percentage of urban area within 100 m of a tubewell
8	Percentage of tree cover within 100 m of a tubewell
9	Average rainfall in 3 days preceding sampling
10	Average temperature in 3 days preceding sampling
11	Average temperature in 15 days preceding sampling
12	Percentage of wetlands within 100 m of a tubewell
13	Average rainfall in 30 days preceding sampling
14	Rainfall on the sampling day
15	Volume of water used for tubewell priming
16	Average temperature in 30 days preceding sampling
17	Population within 200 m of a tubewell
18	Whether a tubewell is a deep well
19	Average temperature in 7 days preceding sampling
20	Population within 100 m of a tubewell
21	Population within 50 m of a tubewell
22	Population within 25 m of a tubewell
23	Tubewell depth
24	Number of hot days in 15 days preceding sampling
25	Distance from a tubewell to the nearest latrine

4. Discussion

The findings of this study provide a comprehensive understanding of the relative importance of various factors influencing fecal contamination in shallow groundwater. By applying machine learning methods, this study not only confirmed previously known relationships but also uncovered new insights into the complex interactions among climatic, land use, and demographic variables. This study is the first to use machine learning techniques to identify and rank influential factors for *E. coli* presence and concentration in shallow tubewell water. The results have implications for developing a forecast system

of groundwater microbial quality and making effective strategies to reduce the risk of waterborne diseases in areas with high groundwater consumption.

E. coli is one of the most commonly used indicators of fecal contamination in drinking water [41] and its presence in groundwater has been frequently reported [11,19,29,42]. In rural Bangladesh, *E. coli* contamination was detected in tubewells more frequently in densely populated areas and during the monsoon season, reflecting that infiltration of surface contaminants into the aquifer is a major pathway for groundwater fecal contamination [29]. Latrine-contaminated ponds are another significant source of fecal contamination in shallow unconfined aquifers [43].

The consistent identification of climatic variables, such as rainfall and temperature, as significant predictors across multiple feature selection methods underscores their pivotal role in influencing fecal contamination in tubewells. Specifically, cumulative rainfall over different time frames (e.g., 30, 15, and 7 days preceding sampling) and short-term climatic events, such as heavy-rain days, emerged as critical contributors. Precipitation was expected to have a positive association with *E. coli* concentration because it can promote fecal contamination of shallow wells [20,29]. Because *E. coli* likely proliferates faster in warm weather than in cold weather, its concentration increases as the temperature rises but then decreases when the temperature is very low. However, high temperatures likely increase the decay of *E. coli* [44]. In addition, high temperatures indicate dry weather, which is likely to influence groundwater recharge.

Land use characteristics, including the percentages of agricultural land, urban areas, tree cover, and barren land within 100 m of a tubewell, were also strongly associated with *E. coli* presence and concentration. The association between *E. coli* concentration and the percentage of urban land and water areas surrounding a tubewell may be due to the high correlation between urban land and population and latrine density, and because pond water is a major source of fecal contamination. Urban and agricultural areas may act as sources of fecal contamination due to the proximity of latrines, livestock operations, and poorly managed wastewater systems.

Some variables, such as population density and the number and type of latrines surrounding the tubewell, were also good predictors of fecal contamination because they might also have considerable influence on *E. coli* concentration in tubewell water, as human waste is a major source of fecal contamination. However, these variables might not be needed in the model because they were highly correlated with land use variables (e.g., the percentage of urban land). Groundwater contamination is often influenced by complex interactions between climatic, land use, and demographic factors. Climatic variables can affect the rate of groundwater recharge and the transport of contaminants through soil layers. Land use patterns contribute to contamination through wastewater discharge from residential areas and animal waste from agricultural lands. Demographic factors, including population density and socioeconomic status, can exacerbate these issues by increasing pressure on local groundwater resources and limiting the availability of sanitation facilities and waste management systems. Understanding these interactions is crucial for developing targeted strategies to mitigate contamination risks and protect groundwater quality.

In these models, some groundwater water quality variables were not included, such as total coliforms, ORP, and pH. These variables might have a closer relationship with fecal contaminants. For example, the association between total coliforms and *E. coli* is obvious because *E. coli* is a subset of bacteria included in total coliform counts, while ORP is a common indicator of water chemistry, which indicates water's ability to receive or gain electrons [45]. However, these water quality variables were excluded from our models due to the high costs associated with data collection. While this may reduce the

model's accuracy in capturing localized contamination dynamics, it broadens the model's applicability to regions where such data are unavailable or unaffordable.

This study used three machine learning methods to identify and rank influential factors. Each offers distinct advantages and limitations. RFE with XGBoost can reduce features by iteratively eliminating less important variables, making it particularly effective for data with a large number of variables. However, its computational intensity and sensitivity to hyperparameter settings may limit its scalability for larger datasets. Random Forest, on the other hand, provides robust importance rankings by averaging over multiple decision trees, offering resilience against overfitting and strong interpretability. Its weakness lies in potential bias toward variables with more levels or variability, which may lead to inflated importance rankings for certain features. Finally, the mutual information uniquely captures non-linear dependencies and interactions between features and the target variable, making it valuable for identifying subtle relationships often missed by tree-based methods. Nonetheless, it does not inherently account for multicollinearity among predictors, which can obscure the independent effects of highly correlated variables. By combining these methods, this study leveraged their complementary strengths to ensure a more comprehensive and nuanced identification of influential factors, providing robust and actionable insights for groundwater quality management.

Given that climatic factors emerged as significant influences on groundwater contamination, future research directions could assess the potential impacts of long-term changes in temperature, precipitation patterns, and extreme weather events on groundwater quality. For instance, changes in rainfall intensity and frequency could influence surface runoff, recharge rates, and contamination pathways, while rising temperatures might affect microbial survival and growth in groundwater. By identifying significant climatic, land use, and demographic variables, future predictive models can leverage these factors to predict contamination risks under various scenarios. In terms of water quality management, such predictive models could guide targeted monitoring efforts, prioritize resource allocation, and inform intervention strategies to mitigate contamination risks. Additionally, these models can help identify vulnerable communities and predict potential outbreaks of waterborne diseases, enabling proactive measures to protect population health.

5. Conclusions

This study identified and ranked key factors potentially influencing fecal contamination in shallow tubewells water using machine learning. The main conclusions are drawn below:

- Climatic variables, land use, and demographic factors were identified as key predictors of fecal contamination in shallow tubewell water, with factors such as rainfall, temperature, land use within 100 m of a tubewell, and population density significantly influencing *E. coli* presence and concentration.
- By identifying influential factors, robust machine learning models can be developed to predict groundwater contamination and prioritize mitigation efforts, providing a data-driven framework for targeted interventions such as land use management and adaptation to climatic variability to improve water quality and public health.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; formal analysis, J.W. and Y.C.; investigation, J.W. and Y.C.; resources, M.S.I. and M.E.; data curation, J.W.; writing—original draft preparation, J.W. and Y.C.; writing—review and editing, J.W., Y.C., M.S.I. and M.E.; visualization, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The data collection was partly supported by grant 5 R01 TW008066 from the United States National Institutes of Health, the Fogarty International Center, and National Science Foundation grant BCS-1560970. The authors also thank core donors, which provide unrestricted support to icddr,b for its operations and research. Current core donors providing unrestricted support to icddr,b include the Governments of Bangladesh, Canada.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Ferrer, N.; Folch, A.; Masó, G.; Sanchez, S.; Sanchez-Vila, X. What are the main factors influencing the presence of faecal bacteria pollution in groundwater systems in developing countries? *J. Contam. Hydrol.* **2020**, *228*, 103556. [[CrossRef](#)]
- Jasechko, S.; Perrone, D. Global groundwater wells at risk of running dry. *Science* **2021**, *372*, 418–421. [[CrossRef](#)] [[PubMed](#)]
- Lall, U.; Josset, L.; Russo, T. A snapshot of the world's groundwater challenges. *Annu. Rev. Environ. Resour.* **2020**, *45*, 171–194. [[CrossRef](#)]
- Murphy, H.M.; Prioleau, M.D.; Borchardt, M.A.; Hynds, P.D. Epidemiological evidence of groundwater contribution to global enteric disease, 1948–2015. *Hydrogeol. J.* **2017**, *25*, 981–1001. [[CrossRef](#)]
- Conboy, M.; Goss, M. Natural protection of groundwater against bacteria of fecal origin. *J. Contam. Hydrol.* **2000**, *43*, 1–24. [[CrossRef](#)]
- Schmoll, O. *Protecting Groundwater for Health: Managing the Quality of Drinking-Water Sources*; World Health Organization: Geneva, Switzerland, 2006.
- Abanyie, S.K.; Apea, O.B.; Abagale, S.A.; Amuah, E.E.Y.; Sunkari, E.D. Sources and factors influencing groundwater quality and associated health implications: A review. *Emerg. Contam.* **2023**, *9*, 100207. [[CrossRef](#)]
- El-Magd, S.A.A.; Ahmed, H.; Pham, Q.B.; Linh, N.T.T.; Anh, D.T.; Elkhrachy, I.; Masoud, A.M. Possible factors driving groundwater quality and its vulnerability to land use, floods, and droughts using hydrochemical analysis and GIS approaches. *Water* **2022**, *14*, 4073. [[CrossRef](#)]
- Alberti, L.; Antelmi, M.; Oberto, G.; La Licata, I.; Mazzon, P. Evaluation of fresh groundwater Lens Volume and its possible use in Nauru island. *Water* **2022**, *14*, 3201. [[CrossRef](#)]
- Keswick, B.H.; Gerba, C.P. Viruses in groundwater. *Environ. Sci. Technol.* **1980**, *14*, 1290–1297. [[CrossRef](#)]
- Macler, B.A.; Merkle, J.C. Current knowledge on groundwater microbial pathogens and their control. *Hydrogeol. J.* **2000**, *8*, 29–40. [[CrossRef](#)]
- Pedley, S.; Howard, G. The public health implications of microbiological contamination of groundwater. *Q. J. Eng. Geol. Hydrogeol.* **1997**, *30*, 179–188. [[CrossRef](#)]
- Dong, Y.; Jiang, Z.; Hu, Y.; Jiang, Y.; Tong, L.; Yu, Y.; Cheng, J.; He, Y.; Shi, J.; Wang, Y. Pathogen contamination of groundwater systems and health risks. *Crit. Rev. Environ. Sci. Technol.* **2024**, *54*, 267–289. [[CrossRef](#)]
- Mahagamage, M.; Pathirage, M.; Manage, P.M. Contamination status of *Salmonella* spp., *Shigella* spp. and *Campylobacter* spp. in surface and groundwater of the Kelani River Basin, Sri Lanka. *Water* **2020**, *12*, 2187. [[CrossRef](#)]
- Gallay, A.; De Valk, H.; Cournot, M.; Ladeuil, B.; Hemery, C.; Castor, C.; Bon, F.; Megraud, F.; Le Cann, P.; Desenclos, J. A large multi-pathogen waterborne community outbreak linked to faecal contamination of a groundwater system, France, 2000. *Clin. Microbiol. Infect.* **2006**, *12*, 561–570. [[CrossRef](#)] [[PubMed](#)]
- Bivins, A.; Lowry, S.; Murphy, H.M.; Borchardt, M.; Coyte, R.; Labhassetwar, P.; Brown, J. Waterborne pathogen monitoring in Jaipur, India reveals potential microbial risks of urban groundwater supply. *Npj Clean Water* **2020**, *3*, 35. [[CrossRef](#)]
- Ferguson, A.S.; Layton, A.C.; Mailloux, B.J.; Culligan, P.J.; Williams, D.E.; Smartt, A.E.; Sayler, G.S.; Feighery, J.; McKay, L.D.; Knappett, P.S. Comparison of fecal indicators with pathogenic bacteria and rotavirus in groundwater. *Sci. Total Environ.* **2012**, *431*, 314–322. [[CrossRef](#)]
- Ercumen, A.; Naser, A.M.; Arnold, B.F.; Unicomb, L.; Colford, J.M., Jr.; Luby, S.P. Can sanitary inspection surveys predict risk of microbiological contamination of groundwater sources? Evidence from shallow tubewells in rural Bangladesh. *Am. J. Trop. Med. Hyg.* **2017**, *96*, 561–568. [[CrossRef](#)] [[PubMed](#)]
- Luby, S.; Gupta, S.; Sheikh, M.; Johnston, R.; Ram, P.; Islam, M. Tubewell water quality and predictors of contamination in three flood-prone areas in Bangladesh. *J. Appl. Microbiol.* **2008**, *105*, 1002–1008. [[CrossRef](#)]
- Wu, J.Y.; Yunus, M.; Islam, M.S.; Emch, M. Influence of Climate Extremes and Land Use on Fecal Contamination of Shallow Tubewells in Bangladesh. *Environ. Sci. Technol.* **2016**, *50*, 2669–2676. [[CrossRef](#)]
- Poulin, C.; Peletz, R.; Ercumen, A.; Pickering, A.J.; Marshall, K.; Boehm, A.B.; Khush, R.; Delaire, C. What Environmental Factors Influence the Concentration of Fecal Indicator Bacteria in Groundwater? Insights from Explanatory Modeling in Uganda and Bangladesh. *Environ. Sci. Technol.* **2020**, *54*, 13566–13578. [[CrossRef](#)]

22. Knappett, P.S.; McKay, L.D.; Layton, A.; Williams, D.E.; Alam, M.J.; Mailloux, B.J.; Ferguson, A.S.; Culligan, P.J.; Serre, M.L.; Emch, M.; et al. Unsealed tubewells lead to increased fecal contamination of drinking water. *J. Water Health* **2012**, *10*, 565–578. [[CrossRef](#)] [[PubMed](#)]
23. White, K.; Dickson-Anderson, S.; Majury, A.; McDermott, K.; Hynds, P.; Brown, R.S.; Schuster-Wallace, C. Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset. *Water Res.* **2021**, *197*, 117089. [[CrossRef](#)] [[PubMed](#)]
24. Howard, G.; Pedley, S.; Barrett, M.; Nalubega, M.; Johal, K. Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. *Water Res.* **2003**, *37*, 3421–3429. [[CrossRef](#)] [[PubMed](#)]
25. Díaz-Alcaide, S.; Martínez-Santos, P. Mapping fecal pollution in rural groundwater supplies by means of artificial intelligence classifiers. *J. Hydrol.* **2019**, *577*, 124006. [[CrossRef](#)]
26. Gómez-Escalonilla, V.; Montero-González, E.; Díaz-Alcaide, S.; Martín-Loeches, M.; del Rosario, M.R.; Martínez-Santos, P. A machine learning approach to site groundwater contamination monitoring wells. *Appl. Water Sci.* **2024**, *14*, 250. [[CrossRef](#)]
27. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
28. Alam, N.; Ali, T.; Razzaque, A.; Rahman, M.; Zahirul Haq, M.; Saha, S.K.; Ahmed, A.; Sarder, A.M.; Moinuddin Haider, M.; Yunus, M.; et al. Health and demographic surveillance system (HDSS) in Matlab, Bangladesh. *Int. J. Epidemiol.* **2017**, *46*, 809–816. [[CrossRef](#)] [[PubMed](#)]
29. van Geen, A.; Ahmed, K.M.; Akita, Y.; Alam, M.J.; Culligan, P.J.; Emch, M.; Escamilla, V.; Feighery, J.; Ferguson, A.S.; Knappett, P.; et al. Fecal Contamination of Shallow Tubewells in Bangladesh Inversely Related to Arsenic. *Environ. Sci. Technol.* **2011**, *45*, 1199–1205. [[CrossRef](#)]
30. Escamilla, V.; Knappett, P.S.; Yunus, M.; Streatfield, P.; Emch, M. Influence of latrine proximity and type on tubewell water quality and diarrheal disease in Bangladesh. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 299–308. [[CrossRef](#)]
31. Wu, J.; Yunus, M.; Ali, M.; Escamilla, V.; Emch, M. Influences of heatwave, rainfall, and tree cover on cholera in Bangladesh. *Environ. Int.* **2018**, *120*, 304–311. [[CrossRef](#)] [[PubMed](#)]
32. Chen, X.-w.; Jeong, J.C. Enhanced recursive feature elimination. In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 13–15 December 2007; pp. 429–435.
33. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
34. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
35. Venkateswarlu, T.; Anmala, J. Importance of land use factors in the prediction of water quality of the Upper Green River watershed, Kentucky, USA, using random forest. *Environ. Dev. Sustain.* **2024**, *26*, 23961–23984. [[CrossRef](#)]
36. Wu, J.; Song, C.; Dubinsky, E.A.; Stewart, J.R. Tracking major sources of water contamination using machine learning. *Front. Microbiol.* **2021**, *11*, 616692. [[CrossRef](#)]
37. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [[CrossRef](#)]
38. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [[CrossRef](#)]
39. Zhou, H.; Wang, X.; Zhang, Y. Feature selection based on weighted conditional mutual information. *Appl. Comput. Inform.* **2024**, *20*, 55–68. [[CrossRef](#)]
40. Kraskov, A.; Stogbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E—Stat. Nonlinear Soft Matter Phys.* **2004**, *69*, 066138. [[CrossRef](#)]
41. Wu, J.; Long, S.C.; Das, D.; Dorner, S.M. Are microbial indicators and pathogens correlated? A statistical analysis of 40 years of research. *J. Water Health* **2011**, *9*, 265–278. [[CrossRef](#)]
42. Atherholt, T.; Feerst, E.; Hovendon, B.; Kwak, J.; Rosen, J.D. Evaluation of indicators of fecal contamination in groundwater. *J. Am. Water Work. Assoc.* **2003**, *95*, 119–131. [[CrossRef](#)]
43. Knappett, P.S.K.; Escamilla, V.; Layton, A.; McKay, L.D.; Emch, M.; Williams, D.E.; Huq, R.; Alam, J.; Farhana, L.; Mailloux, B.J.; et al. Impact of population and latrines on fecal contamination of ponds in rural Bangladesh. *Sci. Total Environ.* **2011**, *409*, 3174–3182. [[CrossRef](#)]
44. Blaustein, R.; Pachepsky, Y.; Hill, R.; Shelton, D.; Whelan, G. Escherichia coli survival in waters: Temperature dependence. *Water Res.* **2013**, *47*, 569–578. [[CrossRef](#)]
45. Hounslow, A. *Water Quality Data: Analysis and Interpretation*; CRC Press: Boca Raton, FL, USA, 1995.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.