Article

Not peer-reviewed version

Predicting Predisposition to Tropical Diseases in Female Adults Using Risk Factors: An Explainable-Machine Learning Approach

Kingsley Friday Attai * , Constance Amannah , Moses Ekpenyong , Said Baadel , Okure Obot , Daniel Asuquo , Ekerette Attai , Faith-Valentine Uzoka , Emem Dan , Christie Akwaowo , Faith-Michael Uzoka *

Posted Date: 22 April 2025

doi: 10.20944/preprints202504.1897.v1

Keywords: Febrile Disease, Explainability, Interpretability, LIME, Machine Learning, Malaria, Random Forest, RTI, Tropical Disease, Typhoid fever, UTI, XAI, XGBoost.

Article

Predicting Predisposition to Tropical Diseases in Female Adults Using Risk Factors: An Explainable-Machine Learning Approach

Kingsley Attai [1,2,*], Constance Amannah [3], Moses Ekpenyong [4,5], Said Baadel [6], Okure Obot [4], Daniel Asuquo [7], Ekerette Attai [2], Faith-Valentine Uzoka [8], Emem Dan [9], Christie Akwaowo [9] and Faith-Michael Uzoka [6]

[1] Department of Mathematics and Computer Science, Ritman University, Ikot Ekpene, Nigeria
[2] Novena Computers and Technologies Limited, Uyo, Nigeria;
[3] Department of Computer Science, Ignatius Ajuru University of Education, Port Harcourt, Nigeria;
[4] Department of Computer Science, Faculty of Computing, University of Uyo, Uyo, Nigeria;
[5] STEM Centre, University of Uyo, and Centre for Research, University of Uyo
[6] Department of Mathematics and Computing, Mount Royal University, Calgary, AB T3E 6K6, Canada;
[7] Department of Information Systems, Faculty of Computing, University of Uyo, Uyo, Nigeria;
[8] Texas Southern University, 3100 Cleburne St, Houston, TX, United States;
[9] Institute of Health Research and Development, University of Uyo Teaching Hospital, Uyo, Nigeria;
* Correspondence: attai.kingsley@ritmanuniversity.edu.ng or attaiekerette@gmail.com; Tel.: +2348101250218

Abstract: Malaria, typhoid fever, respiratory tract infections, and urinary tract infections significantly impact women, especially in remote, resource-constrained settings, due to limited access to quality healthcare and certain risk factors. Most studies have focused on vector control measures, such as insecticide-treated nets and time series analysis, often neglecting emerging yet critical risk factors vital for effectively preventing febrile diseases. We address this gap by investigating the use of machine learning (ML) models, specifically Extreme Gradient Boost and Random Forest, in predicting adult females' susceptibility to these diseases based on biological, environmental, and socioeconomic factors. Explainable AI (XAI) techniques, such as Local Interpretable Model-Agnostic Explanations (LIME), were applied to enhance the transparency and interpretability of these models. This approach provided insights into the models' decision-making process and identified key risk factors, enabling healthcare professionals to personalize treatment services. Factors such as high cholesterol levels, poor personal hygiene, and exposure to air pollution emerged as significant contributors to disease susceptibility, revealing critical areas for public health intervention in remote and resource-constrained settings. This study demonstrates the effectiveness of integrating XAI with ML in directing health interventions, providing a clearer understanding of risk factors, and efficiently allocating resources for disease prevention and treatment.

Keywords: Febrile Disease; Explainability; Interpretability; LIME; Machine Learning; Malaria; Random Forest; RTI; Tropical Disease; Typhoid fever; UTI; XAI; XGBoost

1. Introduction

Tropical diseases such as urinary tract infection (UTI), respiratory tract infection (RTI), malaria, and typhoid fever are significant health concerns. In low- and medium-income countries (LMICs), these diseases impact the vulnerable groups, especially the female population, due to non-clinical risk factors categorized as environmental, socioeconomic, and biological factors [1]. Environmental factors are outside elements about a person's physical surroundings and living circumstances that can potentially raise their risk of developing illnesses. Poor sanitation, overcrowded living spaces, exposure to mosquitoes, and traveling to regions where tropical diseases are endemic are all environmental factors that increase disease susceptibility. Direct contact with infected persons as well as pollution of the environment are also important risk factors in transmitting the disease. Socioeconomic factors constitute an individual's financial, occupational, and social circumstances and can impact their health and ability to access healthcare services. For example, street vendors may experience poor personal hygiene or live in unsanitary conditions, elevating their risk of disease. Poor access to clean water, malnutrition, intravenous drug use, and limited healthcare resources can further contribute to disease susceptibility. Biological risk factors are genetic or physical characteristics of an individual, including pre-existing health conditions such as high blood pressure, high cholesterol, underlying chronic illnesses, and genetic predispositions, which can compromise immunity and increase susceptibility to infections. The heightened susceptibility of an individual to tropical diseases can also be attributed to allergies and other biological vulnerabilities that impact the host's immune system.

Women, particularly in LMICs, are affected by tropical diseases because of a confluence of environmental, socioeconomic, and biological factors. Their vulnerability to infections such as UTIs and malaria is heightened by biological factors such as hormonal fluctuations, pregnancy, as well as anatomical variations. Hormonal fluctuations during the menstrual cycle, pregnancy, and menopause can impact the immune system, which could lead to heightened vulnerability to infections [2–4]. The increased risk of UTI in women is attributed to the relatively short female urethra. This results in a reduction in the distance at which bacteria such as Enterococcus fecalis, Streptococcus species, and Escherichia coli move from the anus into the urethra. In addition, the female urethra opens up into the vulvar vestibular (which is prone to frequent vaginal infections); thus, during sexual activity and when using female hygiene products, the balance of the natural microbacteria of the vagina is distorted [5,6]. Due to increasing numbers of caesarean section, perioperative catheterisations, as well as vaginal examinations during labour, urinary tract infections are typically common during pregnancy and the perinatal period. On the other hand, the post-menopausal woman's risk increases due to a fall in estrogen and glycogen levels leading to vaginal epithelial atrophy as well as a reduction in lactic acid bacteria counts. This leads to the spread and infection of the urinary tract by other bacteria, primarily Escherichia coli. According to estimates, between 10 and 60 percent of women will at some point in their lives get a symptomatic UTI, and every other woman will have experienced at least one UTI [7,8]. UTIs are more common in women but are more severe in men [9]. Temporary suppression of the immune system is a characteristic of pregnancy that enables the body to strike a balance between shielding the foetus from the mother's immune system and shielding the mother from infection [10]. In Mehta & Mann [11], this balance is seen in the numerous physiological, immunological, and anatomical changes that take place to accommodate the developing foetus. These changes can also increase the severity of some infections and make pregnant women more vulnerable.

Women are also at risk of respiratory tract infection and hospitalization, especially pregnant women [12,13]. A twofold higher risk of being overweight or obese in the female gender has been reported, thus increasing the risk of obesity-related physical and psychological comorbidities [14]. Obesity increases the risk of respiratory tract infections, and thus, women have a higher risk of respiratory tract infections [15]. Socially, childcare, caring for ailing family members, and managing household chores are more frequently performed by women in LMICs, including cooking and cleaning, often in unsanitary conditions, which puts them in closer contact with contaminated areas. Since women are the primary caregivers in most families, they spend more time with ailing relatives [16], which increases their susceptibility to diseases. These environments increase their exposure to contaminated food and water, increasing their risk of contracting infections. They also frequently spend time indoors using biomass fuels for cooking, increasing their risk of respiratory tract infections from exposure to smoke and other pollutants [17]. Furthermore, their economic roles, such as working in agriculture or street vending, increase their exposure to vectors like mosquitoes. In resource-poor settings, a large number of women work outside in jobs as street vendors [18], agriculture [19], and construction [20], which increases their exposure to mosquitoes and contaminated environments, thereby raising their risk of contracting UTI, typhoid fever, and malaria.

There is also gender disparity in literacy levels, with men being more literate than women [21]. Thus, knowledge about these disease conditions, hygiene as well and prevention is limited in women compared to men. Women are more likely to be chronic carriers of Typhoid [22], which increases their risk of re-infection, thus putting them more at risk of infections. The abovementioned risk factors highlight the pressing need for research on disease susceptibility to concentrate on the female population. By developing predictive models that target this population, public health policies can be improved, particularly in areas with limited medical resources and health disparities, thereby improving prevention and treatment strategies. The prevention of severe health outcomes and the effectiveness of treatment depend on the early and accurate detection of tropical diseases. Conventional machine learning techniques have been applied in predicting tropical conditions, using patient symptoms and other clinical features.

Most studies have applied ML and XAI techniques with clinical features such as symptoms, laboratory findings, and pathogen characteristics, to diagnose typhoid fever, malaria [23–25], UTI [26], and RTI [27], but there seems to be a gap in addressing how non-clinical factors particularly environmental, socioeconomic, and biological risk factors have in determining an individual's susceptibility to tropical disease, particularly in vulnerable groups such as women. Our study addresses this gap by shifting the focus to non-clinical factors and how these risk factors increase the susceptibility to diseases like typhoid, malaria, respiratory, and urinary infections. This novel approach complements clinical studies, providing a comprehensive understanding of disease risk that is vital for preventive measures, particularly in resource-poor settings where clinical interventions may be delayed or inaccessible. The predictive models developed in this study hold great potential for personalized interventions and public health strategies targeted at lowering the incidence of tropical diseases among women. Through the identification of women who are more susceptible to health risks due to these factors, policymakers and healthcare professionals can better allocate resources, concentrating on preventive measures like vector control, better sanitation, and health education in communities at high risk. Personalized interventions, in which women who are considered to be at risk receive specific guidance, immunizations, or medical care, can also be made easier by these predictions, guaranteeing that public health initiatives are both focused and economical. Predicting susceptibility aids in the distribution of scarce medical resources and guarantees that the most vulnerable groups receive treatments and supplies on time.

The advent of Explainable Artificial Intelligence (XAI) methods like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) offers transparency into the decision-making process. The XAI methods facilitate transparency by providing insights into the decision-making processes of intricate machine learning models [28,29]. Since ML models, which are sometimes referred to as "black boxes," generate accurate predictions that are not interpretable, it makes it challenging for users in the health sector to comprehend and trust their outcomes. To address this problem, LIME builds locally interpretable models around individual predictions, highlighting the key features (risk factors) influencing a prediction and explaining why specific risk factors are involved in specific outcomes. This increases user trust and accountability and improves the transparency and dependability of AI-driven systems. By incorporating LIME for explainability, clinicians, policymakers, and other stakeholders can have greater confidence in the transparent decision-making process of the ML models. Healthcare practitioners can better understand why some women are considered more susceptible to particular tropical diseases by using these explainability techniques, which provide a thorough breakdown of the risk factors that most strongly influence each prediction. To guarantee that AI-driven insights are applicable and enable better-informed decision-making and focused interventions, this transparency is essential. It provides a mechanism for policymakers to rank risk mitigation initiatives according to distinct, comprehensible criteria, encouraging evidence-based policies targeted at lowering the burden of disease in high-risk populations.

This study aims to use LIME and ML models to predict the susceptibility of women to tropical diseases based on environmental, socioeconomic, and biological risk factors. By concentrating on this susceptible population, we hope to develop a predictive framework that improves prediction accuracy and healthcare outcomes for populations at high risk of tropical diseases by integrating ML and XAI as demonstrated in this work. Some primary contributions of this study are:

- By using ML algorithms such as Random Forest (RF) and Extreme Gradient Boost (XGBoost), the study improves the prediction of susceptibility to tropical diseases and offers a data-driven approach to disease prevention and intervention by efficiently processing risk factors to provide accurate predictions for at-risk women.
- The integration of LIME with XGBoost and RF offers explainability for the model's predictions, making decisions more comprehensible and practical by enabling policymakers and healthcare professionals to comprehend the precise risk factors influencing each prediction. This transparency builds trust in AI-driven healthcare solutions and enables targeted interventions based on identified risk factors.

The study is structured as follows: Section 2 outlines the methodology, which includes data collection, preprocessing, the suggested system framework, the prediction and interpretability model for enhanced diagnostic interpretability, and the model's performance metrics. Section 3 presents the findings and discussion, assessing the performance of different algorithms and demonstrating how XAI provides information about model choices. The study is concluded in Section 4, which also identifies its shortcomings and suggests additional research.

2. Methodology

2.1. Dataset Description and Data Preprocessing

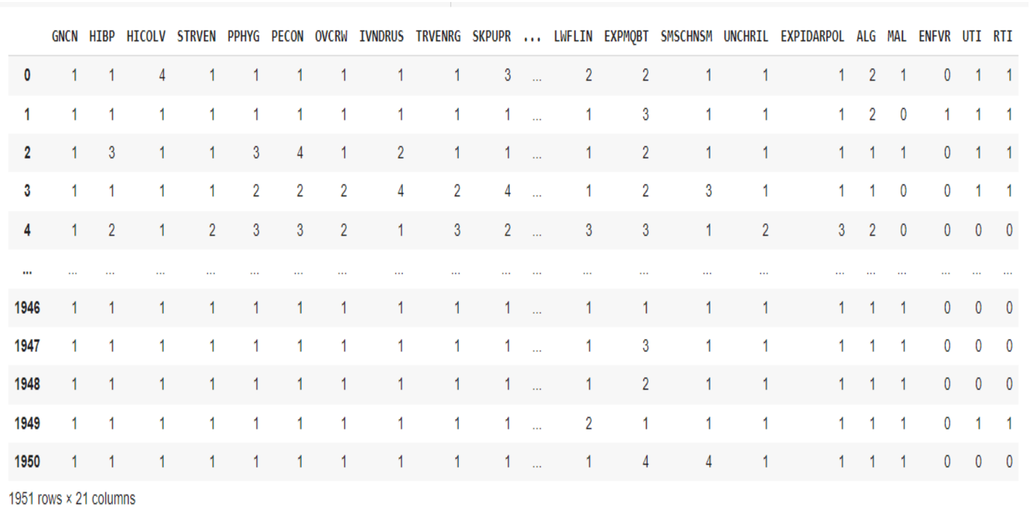
The New Frontiers in Research Fund (NFRF) project provided 4870 patient records for this study [30]. The dataset was segmented according to patient symptoms, demographic data, risk factors, suspected diagnosis, additional testing, and confirmed diagnosis. Five points were assigned to each patient's symptom severity and risk factors on the dataset: 5 for very severe, 4 for severe, 3 for moderate, 2 for mild, and 1 for absent. The medical professionals made suspected diagnoses based on the extent to which the patient is susceptible to non-clinical risk factors. Before reporting the confirmed diagnoses, additional tests, including blood films, serology, complete blood counts, etc., were performed. Using a language scale, the severity of the suspected and confirmed diagnoses was rated: 6 = very high; 5 = high; 4 = moderate; 3 = low; 2 = very low; and 1 = absent. Malaria, HIV/AIDS, TB, typhoid fever, dengue fever, urinary tract infection, yellow fever, respiratory tract infection, and Lassa fever were the illnesses included in the dataset. To preserve the integrity of the dataset, records with missing symptoms, risk factors, and diseases (TB, HIV/AIDS, dengue fever, yellow fever, and Lassa fever) not covered by this study were removed during data preprocessing.

Women from adolescents (13 years to 18 years) and above were chosen for this study. Adolescents go through a crucial period of physical, hormonal, and immune system changes. Adolescent women frequently undergo biological changes during this time, including the onset of menstruation and hormonal fluctuations, which can impact their vulnerability to infections. Furthermore, the socioeconomic difficulties that many teenagers in LMICs face, such as poor living conditions, poor nutrition, and restricted access to health care, increase their susceptibility. By including adolescents, this study captures a broad spectrum of life stages where women are susceptible to non-clinical risk factors, providing comprehensive insights into how environmental, socioeconomic, and biological influences affect disease susceptibility throughout different phases of adulthood. This ensures that public health interventions can be tailored to address the unique needs of both younger and older women, ultimately improving healthcare outcomes for this vulnerable population. The demographic data of the female patients extracted for this study is presented in Table 1. The age range is presented in five (5) groups (adolescents, young adults, middle-aged adults, older adults, and elderly), as well as the number of pregnant patients from the first to third trimester and nursing mothers.

Table 1. Demographic Data of female patients.

Age range	**Frequency**
13 years to 18 years	182
19 years to 35 years	978
36 years to 50 years	425
51 years to 65 years	260
66 years and above	106
Total	1951
Pregnant Patients	**Frequency**
0-3months	135
4-6months	184
7-9months	86
Total	405
Nursing mothers	**Frequency**
0-3months	26
4-6months	35
7-9months	28
over 9months	61
Total	150

After removing the columns for symptoms, further testing, and doctors' suspected diagnoses, the dataset was narrowed down to risk factors and confirmed diagnoses that fell within the purview of the investigation. By using binary encoding to convert the output labels to absent (0) [1=absent] and present (1) [6 = very high; 5 = high; 4 = moderate; 3 = low and 2 = very low], the classification tasks were made simpler, the model's complexity was decreased, and performance was enhanced because there would be no need for the model to differentiate between multiple, potentially overlapping classes. Only 1951 records remained in the dataset after preprocessing, as Fig. 1 illustrates, with 17 risk factors and 4 confirmed diagnoses. The list of risk factors and the diseases considered in this study is presented in Table 2.



	GNCN	HIBP	HICOLV	STRVEN	PPHYG	PECON	OVCRW	IVNDRUS	TRVENRG	SKPUPR	...	LWFLIN	EXPMQBT	SMSCHNSM	UNCHRIL	EXPIDARPOL	ALG	MAL	ENFVR	UTI	RTI
0	1	1	4	1	1	1	1	1	1	3	...	2	2	1	1	1	2	1	0	1	1
1	1	1	1	1	1	1	1	1	1	1	...	1	3	1	1	1	2	0	1	1	1
2	1	3	1	1	3	4	1	2	1	1	...	1	2	1	1	1	1	1	0	1	1
3	1	1	1	1	2	2	2	4	2	4	...	1	2	3	1	1	1	0	0	1	1
4	1	2	1	2	3	3	2	1	3	2	...	3	3	1	2	3	2	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1946	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1	1	1	0	0	0
1947	1	1	1	1	1	1	1	1	1	1	...	1	3	1	1	1	1	1	0	0	0
1948	1	1	1	1	1	1	1	1	1	1	...	1	2	1	1	1	1	1	0	0	0
1949	1	1	1	1	1	1	1	1	1	1	...	2	1	1	1	1	1	1	0	1	1
1950	1	1	1	1	1	1	1	1	1	1	...	1	4	4	1	1	1	1	0	0	0

1951 rows × 21 columns

Figure 1. Pre-processed dataset.

Table 2. Risk factors and Diseases with abbreviations.

Biological Factors	**Abbreviation**
Poor Environmental Condition	PECON
Overcrowding	OVCRW
Travel to Endemic Region	TRVENRG
Exposure to Mosquito Bite	EXPMQBT
Indoor Air Pollution	EXPIDARPOL
Smoking Exposure	SMSCHNSM
Contact with an Infected Person	DRCOIFPS
Skin Puncture	SKPUPR
Socioeconomic Factors	
Street Vendor	STRVEN
Poor Personal Hygiene	PPHYG
Intravenous Drug Use	IVNDRUS
Low Fluid Intake	LWFLIN
Biological Factors	
Genetic Condition	GNCN
High Blood Pressure	HIBP
High Cholesterol Level	HICOLV
Underlying Chronic Illness	UNCHRIL
Allergy	ALG
Diseases	
Malaria	MAL
Enteric fever (Typhoid Fever)	ENFVR
Urinary Tract Infection	UTI
Respiratory Tract Infection	RTI

2.2. Prediction and Interpretability Models

Google Colab was used in the study along with matplotlib, sk-learn, numpy, and pandas, among other Core Python libraries and packages. The RF and XGBoost algorithms created the prediction models with performance metrics. GridSearchCV, a hyperparameter tuning technique, was included to increase the prediction accuracy. The hyperparameter settings used were RF ('max_depth': [None, 10, 20], 'n_estimators': [100,200,300]) and XGBoost ('max_depth': [3,5,7], 'n_estimators': [100,200,300]). While n_estimators indicates the number of trees to be built, max_depth specifies the maximum depth of the trees. These hyperparameters help to fine-tune the model's behaviour, improving its functionality and capacity to accurately and broadly identify the febrile conditions that this study is considering. The corresponding machine learning algorithms' built-in features were used to derive these hyperparameters. The Random Forest and XGBoost algorithms were utilized in this study. To increase prediction accuracy, the RF algorithm is an ensemble machine learning technique that combines several decision trees with a strong resistance to over-fitting [31]. RF efficiently handles high-dimensional and complex problems and performs well with large datasets [32]. Voting on individual tree predictions creates the final prediction, lowering overfitting and increasing the model's resilience. The gradient boosting framework includes the XGBoost algorithm, which can be applied to resolve problems with predictive modeling for regression or classification. With each new learner focusing on correcting the errors made by the more experienced ones, XGBoost brings in weaker students to the group. XGBoost has gained widespread use in numerical applications, such as illness prediction, due to its well-known ability to manage structured data [33]. LIME uses an interpretable model to approximate the complex model near a specific prediction to provide local explanations [25,34]. This made it possible to produce a graphic explanation that illustrates how the characteristics of the risk factors influenced the forecasts. It indicates which symptoms influenced the model's judgment the most, making the reasoning behind it transparent for healthcare professionals. It also offers succinct, locally interpretable explanations for each prediction.

2.3. Proposed System Framework

The proposed tropical disease prediction framework is presented in Fig. 2. The key components of the framework, which facilitate decision-making, are medical experts, a mobile device for processing and storing data locally and in the cloud, and a healthcare worker. Medical professionals provided the patient data, which was then preprocessed into a format that could be used for ML and XAI modeling. Preprocessing helps the model produce more reliable predictions by ensuring data quality, choosing and encoding relevant features, balancing the dataset, and normalizing inputs. Using a mobile device, a healthcare professional can use the suggested model to diagnose tropical diseases with greater accuracy and understanding. Healthcare professionals can enter a patient's vitals and risk factors using sliders and drop-down menus on the user-friendly interface. Following data entry, the model can instantly process the information and determine whether the patient is most likely to have malaria, typhoid fever, UTI, or RTI.

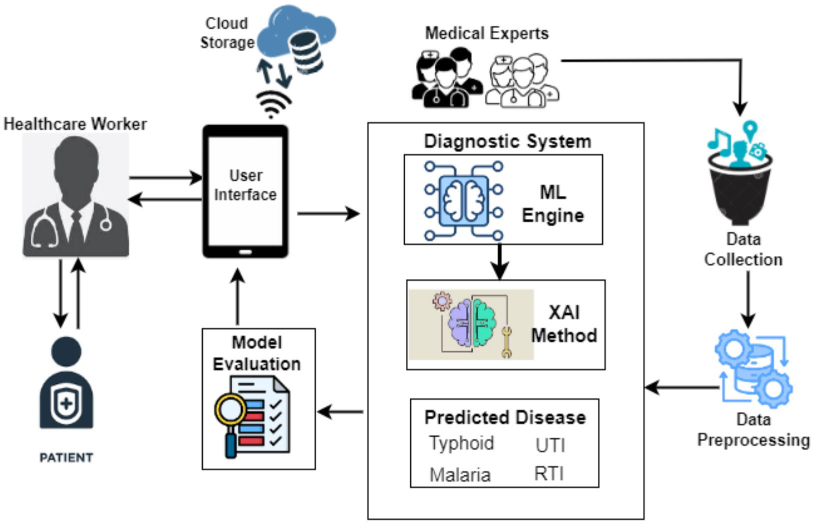


Figure 2. Tropical Disease Prediction Framework.

doi:10.20944/preprints202504.1897.v1

2.4. Model Performance Metrics

Initial records for 4870 patients with feverish symptoms made up the dataset used for this study. Following preprocessing, the number of patient records was reduced to 1951, and only those with pertinent features were kept for machine learning modeling. The dataset was split into 20% for testing and 80% for training. To achieve robust and objective results, for cross-validation, StratifiedKFold was used, rearranging the data before dividing it into five stratified folds. GridSearchCV was used to optimize model performance. Key performance metric components were used to assess the experimental models.

Precision quantifies how well a model detects true positive cases while avoiding false positives. It conveys the precision with which a model forecasts favourable results. Accuracy is important when false positives have a significant cost. For example, a false positive in a medical diagnosis could lead to unnecessary treatments.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The ability of a model to recognize each positive instance in a dataset is measured using a metric known as recall. It measures how sensitive the model is to True Positive cases. In medical screenings, where it can be crucial to miss a positive case (false negative), recall is crucial when the cost of false negatives is high.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The harmonic mean of recall and precision is represented by the F1-Score. The F1-score can have a range of binary values, with 1 representing each class's correctly predicted data point and 0 representing any class's incorrectly predicted data point. The F1 Score can be useful when you need to balance recall and precision, especially if your class distribution is not uniform.

$$F1 = 2 * \left(\frac{Precision\ *\ Recall}{recision +\ Recall}\right)$$

The need to balance accuracy and reliability in a real-world healthcare setting led to the decision to use precision, recall, and F1 score to evaluate the models. These metrics are particularly well-suited for the prediction of tropical diseases, where an overdiagnosis or underdiagnosis can have major implications on women's health, resource allocation, and public health strategies.

3. Results and Discussion

This section displays the results of the evaluation of the models' performance as well as the XAI technique used to predict tropical diseases based on risk factors. Table 2 displays each model's performance, and the model's performance according to the metrics considered is displayed in Fig. 3. XGBoost performs well in malaria prediction, with high recall and precision rates. While recall shows that 84% of those at risk are correctly identified, precision indicates that 89% of the patients identified as susceptible to malaria are actually at risk. With an F1-score of 86%, this model is deemed reliable for predicting susceptibility to malaria, as it demonstrates a balanced trade-off between precision and recall. Typhoid fever is predicted by the model with a moderate degree of precision (64%) but a relatively low recall rate (34%). The F1-score of 44% indicates a significant imbalance, signifying that the model has difficulty correctly identifying women who are susceptible to typhoid fever. The model's moderate precision of 64% for UTI prediction is accompanied by a very low recall of 26%, and its poor performance is reflected in its F1-score of 37%, suggesting that the model is not effective in capturing a patient's susceptibility to UTIs. The model has a 67% precision and a 32% recall rate, which is considered moderate for predicting RTIs. The model can accurately identify 67% of the predicted at-risk cases, according to the F1-score of 43%; however, it can only identify 32% of the true at-risk cases, resulting in a large number of missed cases.

Similar to XGBoost, RF predicts malaria well, yielding almost identical outcomes. An F1-score of 87% indicates balanced and trustworthy predictions, and the high precision (89%) and recall (85%)

indicate that the model is both accurate in its predictions and effective in identifying those who are truly at risk. For typhoid fever, RF outperforms XGBoost in terms of precision (74%). However, recall is still low at 27%, meaning many susceptible women are missed. The model appears to be having difficulty capturing a significant enough number of true positive cases, as indicated by the F1-score of 40%. While RF predicts UTIs with a higher precision (71%) than XGBoost, it still performs poorly in recall (21%). The model's poor F1-score of 32% indicates a limited ability to accurately predict the risk of UTIs by demonstrating an ineffectiveness at striking a balance between precision and recall. RF outperforms XGBoost for RTIs, achieving a 30% recall and 70% precision. Even with acceptable precision, a sizable portion of those who are actually at risk are still missed by the model. The model's moderate ability to identify women who are susceptible to RTIs is reflected in the F1-score of 42%.

Table 2. Prediction model performance.

		MAL	ENFVR	UTI	RTI
XGBoost	**Precision**	0.89	0.64	0.64	0.67
	Recall	0.84	0.34	0.26	0.32
	F1-score	0.86	0.44	0.37	0.43
RF	**Precision**	0.89	0.74	0.71	0.70
	Recall	0.85	0.27	0.21	0.30
	F1-score	0.87	0.40	0.32	0.42

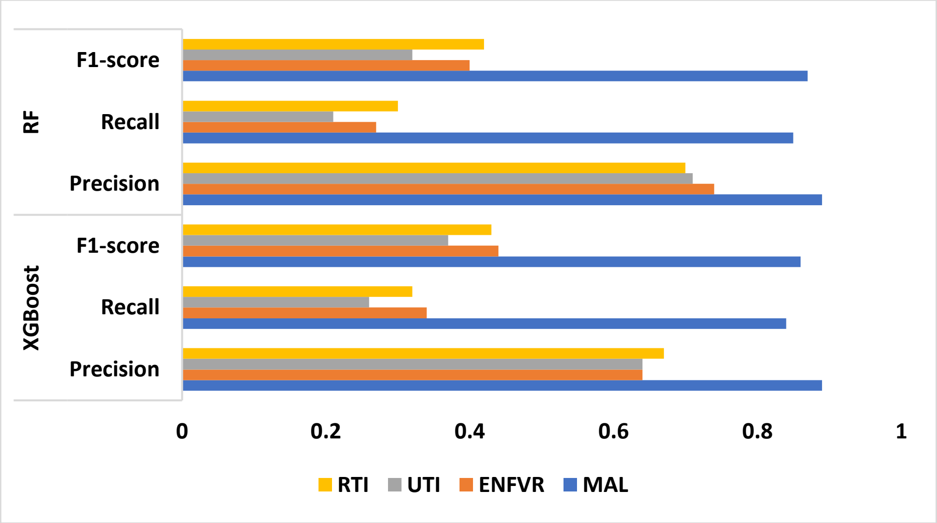


Figure 3. Performance Evaluation of the Models.

These results suggest that while ML models, particularly random forest, do well in predicting malaria, they have trouble with illnesses like respiratory tract infections, typhoid fever, and UTIs. Given the reliability of malaria predictions, patients who have been identified as susceptible can benefit from focused preventive measures like prophylactic treatments or mosquito control. The low recall for illnesses like UTIs and typhoid, however, raises the possibility that many women who are at risk may not be identified, which would reduce the efficacy of public health initiatives aimed at preventing these infections. The model's performance was influenced by the size of the dataset, as both the quantity and quality of data are important considerations in machine learning. This is because smaller datasets can cause overfitting, a condition in which the model performs well on training data but finds it difficult to predict accurately from unseen data. This is demonstrated by the lower recall scores for illnesses such as urinary tract infections and typhoid fever in both Random Forest and XGBoost, which suggests that the model may have several true positive cases, potentially due to insufficient data representing these conditions. Furthermore, imbalances in the way the model learns from the data could result from a smaller dataset's inability to sufficiently capture the diversity of non-clinical risk factors across various demographics and disease profiles. For instance, the lower F1 scores for respiratory tract infections and typhoid fever indicate that there may not have been enough instances of these illnesses in the dataset for the model to properly identify their patterns. By adding more representative data, expanding the dataset may enhance performance by enabling the model to more accurately detect patterns and relationships between non-clinical risk factors and disease susceptibility, particularly in underrepresented categories.

The LIME results for XGBoost in Fig. 4 provide insight into how various risk factors affect the model's predictions for susceptibility to tropical diseases. The negative contribution of mosquito bites and travel to endemic regions seems counterintuitive given the association of these risk factors and diseases like malaria. It might indicate either model overfitting or potential data imbalances in capturing mosquito exposure and travel to endemic regions, suggesting a need for further investigation. The positive contributions of factors like high cholesterol levels, smoking exposure, and poor personal hygiene indicate that these factors can increase the likelihood of susceptibility. While high cholesterol itself isn't directly linked to tropical diseases, this finding may point to underlying health vulnerabilities or general immune system weaknesses that make women more susceptible to infections. This underscores the importance of addressing lifestyle factors and environmental conditions in preventive healthcare strategies for women, especially in regions prone to tropical diseases. Additionally, factors like indoor air pollution and contact with infected persons highlight the need for public health interventions that focus on improving air quality and hygiene education to lower disease transmission risks in vulnerable populations.

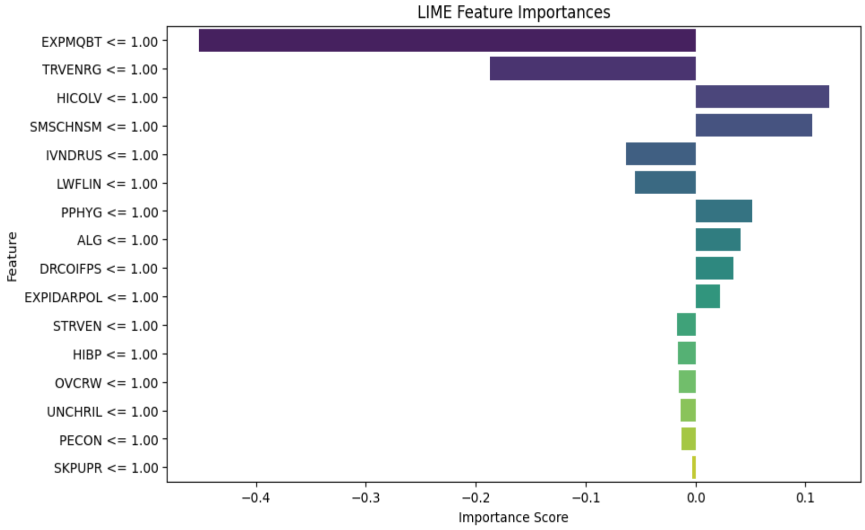


Figure 4. XGBoost Algorithm LIME diagram.

The LIME results for Random Forest reveal key insights into how various risk factors affect the model’s predictions of susceptibility to tropical diseases. The negative influence of traveling to endemic regions is similar to the XGBoost LIME results, which may suggest that women who travel to such regions possibly take preventive measures, leading to a reduced risk of infection. The negative influence of low fluid intake indicates that the dataset doesn't link this risk factor to tropical diseases, although dehydration can sometimes exacerbate disease symptoms. On the other hand, positive contributions of factors such as smoking exposure, indoor air pollution, poor personal hygiene, and contact with infected persons point to areas where public health interventions could play a critical role in reducing disease risk. These findings highlight the need for targeted public health policies that address environmental and lifestyle factors to protect women, particularly in resource-limited settings where tropical diseases are prevalent.

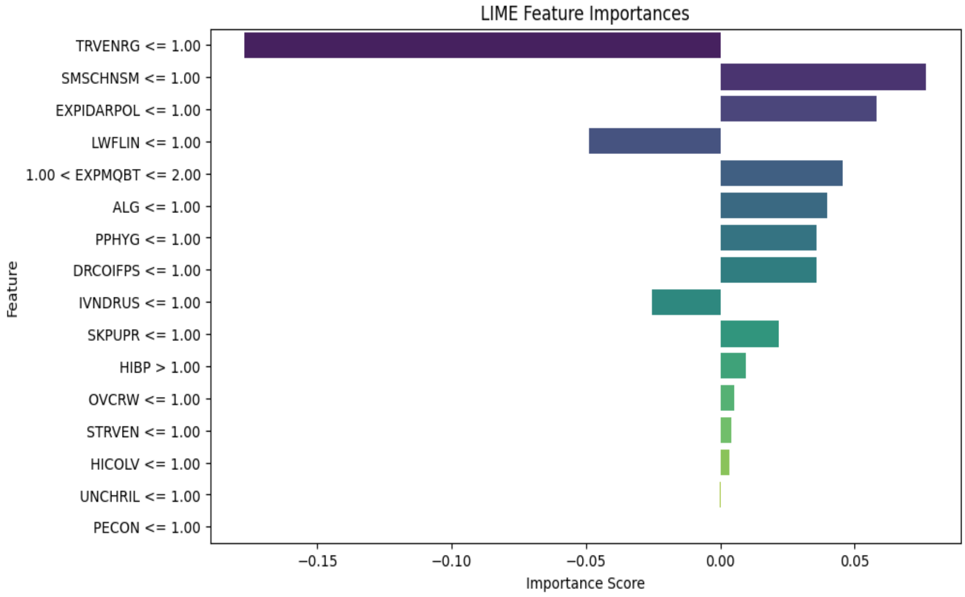


Figure 5. RF Algorithm LIME diagram.

In comparing the LIME results for XGBoost and Random Forest (RF), we see notable similarities and differences in the feature importance rankings and the magnitude of their contributions. Both models show "Travel to Endemic Regions" as an important negative predictor, suggesting that travel may have been mitigated by preventive measures. On the positive side, XGBoost places more emphasis on Smoking Exposure and High Cholesterol, while RF assigns less importance to these and gives more weight to Indoor Air Pollution and Smoking Exposure. Overall, XGBoost's results are more extreme, with higher magnitude feature importance scores, which may indicate a more complex understanding of how certain risk factors affect susceptibility. However, RF's results are more balanced and stable across multiple factors, potentially leading to a more robust and interpretable model. Depending on the context, RF's more consistent and balanced outputs could be considered better for real-world applications, especially when dealing with stakeholders who need interpretable and actionable insights.

4. Conclusions

This study explores the intricate connections between environmental, socioeconomic, and biological risk factors and their influence on the susceptibility of women to tropical diseases such as malaria, typhoid fever, urinary tract infections (UTIs), and respiratory tract infections (RTIs). Women in LMICs are disproportionately affected by these diseases due to their socioeconomic roles, caregiving responsibilities, and increased exposure to environmental hazards. By incorporating non-clinical factors such as poor personal hygiene, overcrowded living conditions, smoking and exposure to smoke, and other socioeconomical factors into predictive models, this research bridges an essential gap in the current understanding of tropical disease susceptibility.

The employed RF and XGBoost models performed well in predicting malaria, making them suitable for preventive measures and targeted interventions. However, the models struggled with diseases like typhoid fever, UTIs, and RTIs, where the recall rates were significantly lower, suggesting overfitting in the training phase. The study also explored the potential of XAI to enhance the interpretability of the models. Incorporation of LIME provided even greater transparency, helping stakeholders better understand the models’ decision-making processes. By using XAI techniques, healthcare professionals and policymakers could understand the key factors influencing the models’ decisions. For example, XAI analysis indicated that factors such as high cholesterol levels, smoking and exposure to smoke, poor personal hygiene, and indoor air pollution were significant contributors to disease susceptibility. This transparency is crucial in ensuring that ML-driven healthcare solutions are trusted and can be integrated into real-world public health strategies.

One major limitation of the study was the reliance on causative risk factors alone. While this approach provided valuable insights into the environmental, socioeconomic, and biological contributors to disease susceptibility, the exclusion of clinical factors such as symptoms and other geospatial and time-series data may have degraded the different models' overall performance in particularly in the prediction of diseases like UTI and RTIs.

Future research should explore the integration of both clinical and causative risk factors to create a more comprehensive predictive framework. This could improve the models' capacity to recognize people who are at risk and improve the overall performance of these febrile disease predictions.

In conclusion, this research offers a fresh perspective on comprehending and predicting disease susceptibility among women in LMICs. By combining advanced ML algorithms with XAI techniques, the research offers valuable insights into the environmental, socioeconomic, and biological factors that contribute to tropical disease susceptibility. The identification of key risk factors, such as exposure to mosquitoes, indoor air pollution, and poor hygiene practices, provides actionable insights for policymakers. By focusing on these areas, public health initiatives can be more effectively tailored to reduce the burden of tropical diseases among women. The findings have significant implications for public health policy, personalized medicine, and the future of AI-driven healthcare.

Author Contributions: Conceptualization, F.-M.U., CH.A., O.O., and K.A.; methodology, K. A., C.A., M.E., D.A. S. B. CH.A., and F.-M.U.; validation, F.-M.U., K. A., M. E., and C.A. .; formal analysis, K. A. D.A., E. A., F.-V.U., E. D.; data curation, K.A., and E.A; writing—original draft preparation, K. A., E. A., S. B., E. D., F.-V.U., CH.A and C.A., and F.-M.U.; writing—review and editing, M.E., D. A., O.O., K. A., C. A., S. B., and F.-M.U. supervision, F.-M.U., M.E., O.O., D.A. and C. A. ; project administration, O.O. and F.-M.U. funding acquisition, F.-M.U. (Constance Amannah=C.A and Christie Akwaowo=CH. A). All authors have read and agreed to the published version of the manuscript

Funding: The New Frontier Research Fund funded this study from April 2020 to March 2024 under grant number NFRFE-2019-01365.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Valdez, G. F. D.; Ajzoon, M.; Al Zuwameri, N. A Scoping Review of the Biological, Socioeconomic and Environmental Determinants of Overweight and Obesity Among Middle Eastern and Northern African Nationalities. Sultan Qaboos Univ. Med. J. 2024, 24 (1), 20. https://doi.org/10.18295/squmj.10.2023.059.
2. Singhal, T. Infections in Pregnancy. J. Clin. Infect. Dis. Soc. 2024, 2 (1), 28–33. https://doi.org/10.4103/CIDS.CIDS_14_24.
3. Jain, J. J. Changing Epidemiology of Infections in Pregnancy: A Global Perspective. In Infections and Pregnancy; Springer Singapore: Singapore, 2022; pp 3–12. https://doi.org/10.1007/978-981-16-7865-3_1.
4. Obeagu, E. I.; Obeagu, O. G. Malaria During Pregnancy: Effects on Maternal Morbidity and Mortality. Elite J. Nurs. Health Sci. 2024, 2 (6), 50–68.
5. Czajkowski, K.; Broś-Konopielko, M.; Teliga-Czajkowska, J. Urinary Tract Infection in Women. Menopause Rev./Przegląd Menopauzalny 2021, 20 (1), 40–47. https://doi.org/10.5114/pm.2021.105382.
6. Faraz, A. A.; Mendem, S.; Swamy, M. V.; Shubham, P.; Vinyas, M. Urinary Tract Infections in Women: Treatment Options and Antibiotic Resistance. J. Pharm. Sci. Res. 2020, 12 (7), 875–879.

7. Curtiss, N.; Meththananda, I.; Duckett, J. Urinary Tract Infection in Obstetrics and Gynaecology. Obstet. Ginecol. Reprod. Med. 2017, 27, 261–265. https://doi.org/10.1016/j.ogrm.2017.06.006.
8. Foxman, B. Urinary Tract Infection Syndromes: Occurrence, Recurrence, Bacteriology, Risk Factors, and Disease Burden. Infect. Dis. Clin. North Am. 2014, 28 (1), 1. https://doi.org/10.1016/j.idc.2013.09.003.
9. Tokatli, M. R.; Sisti, L. G.; Marziali, E.; Nachira, L.; Rossi, M. F.; Amantea, C.; Malorni, W. Hormones and Sex-Specific Medicine in Human Physiopathology. Biomolecules 2022, 12 (3), 413. https://doi.org/10.3390/biom12030413.
10. Kareva, I. Immune Suppression in Pregnancy and Cancer: Parallels and Insights. Transl. Oncol. 2020, 13 (7), 100759. https://doi.org/10.1016/j.tranon.2020.100759.
11. Mehta, S.; Mann, A. Pregnancy Changes Predisposing to Infections. In Infections and Pregnancy; Springer Singapore: Singapore, 2022; pp 13–25. https://doi.org/10.1007/978-981-16-7865-3_2.
12. Dawood, F. S.; Garg, S.; Fink, R. V.; Russell, M. L.; Regan, A. K.; Katz, M. A.; Fell, D. B. Epidemiology and Clinical Outcomes of Hospitalizations for Acute Respiratory or Febrile Illness and Laboratory-Confirmed Influenza among Pregnant Women during Six Influenza Seasons, 2010–2016. J. Infect. Dis. 2020, 221 (10), 1703–1712. https://doi.org/10.1093/infdis/jiz670.
13. Dirican, A. Ö.; Ceran, M. U.; Özçimen, E. E.; Çulha, A. A.; Abasıyanık, M. A.; Üstün, B.; Akgün, S. COVID-19 Infection and Women's Health; Which Women Are More Vulnerable in Terms of Gynecological Health? Preprint. https://doi.org/10.21203/rs.3.rs-3079652/v1.
14. Kapoor, N.; Arora, S.; Kalra, S. Gender Disparities in People Living with Obesity - An Unchartered Territory. J. Midlife Health 2021, 12 (2), 103–107. https://doi.org/10.4103/jmh.jmh_48_21.
15. Maccioni, L.; Weber, S.; Elgizouli, M.; Obesity and Risk of Respiratory Tract Infections: Results of an Infection-Diary-Based Cohort Study. BMC Public Health 2018, 18 (271). https://doi.org/10.1186/s12889-018-5172-8.
16. Asuquo, E. F.; Akpan-Idiok, P. A. The Exceptional Role of Women as Primary Caregivers for People Living with HIV/AIDS in Nigeria, West Africa. Suggestions for Addressing Clinical and Non-Clinical Issues in Palliative Care, Caregiving and Home Care 2020, 101–115. https://doi.org/10.5772/intechopen.93670
17. Zewdie, A.; Degefa, G. H.; Donacho, D. O. Health Risk Assessment of Indoor Air Quality, Sociodemographic and Kitchen Characteristics on Respiratory Health Among Women Responsible for Cooking in Urban Settings of Oromia Region, Ethiopia: A Community-Based Cross-Sectional Study. BMJ Open 2023, 13 (6), e067678. https://doi.org/10.1136/bmjopen-2022-067678.
18. Saad, S. Women and Places; Female Street Vendors, Territorial Identity and Placemaking. J. Art Des. 2022, 1–14. https://doi.org/10.31586/jad.2022.297.
19. Pradhan, S.; Raksha, G. S.; Akhil, P. Role of Women in Food and Agricultural Development: Breaking Barriers for Sustainable Growth. In Impact of Women in Food and Agricultural Development; IGI Global, 2024; pp 130–148. https://doi.org/10.4018/979-8-3693-3037-1.ch008.
20. Statista. Global Adult Literacy Rate from 2000 to 2022, by Gender. https://www.statista.com/statistics/1220131/global-adult-literacy-rate-by-gender/ (accessed 2024).
21. Jimoh, R.; Adamu, A.; Oyewobi, L.; Bajere, P. How Women Are Locked Out of Nigeria’s Construction Industry. The Conversation. Retrieved September 21, 2024, from https://theconversation.com/how-women-are-locked-out-of-nigerias-construction-industry-157643#:~:text=In%20Nigeria%2C%20women%20make%20up,ethics%20and%20values%20in%20Nigeria.

22. Masuet-Aumatell, C.; Atouguia, J. Typhoid Fever Infection – Antibiotic Resistance and Vaccination Strategies: A Narrative Review. Travel Med. Infect. Dis. 2021, 40 (101946). https://doi.org/10.1016/j.tmaid.2020.101946.
23. Muhammad, B.; Varol, A. A Symptom-Based Machine Learning Model for Malaria Diagnosis in Nigeria. In 2021 9th International Symposium on Digital Forensics and Security (ISDFS); IEEE, 2021; pp 1–6. https://doi.org/10.1109/ISDFS52919.2021.9486315.
24. Odion, P. O.; Ogbonnia, E. O. Web-Based Diagnosis of Typhoid and Malaria Using Machine Learning. Nigerian Defence Academy J. Military Sci. Interdiscip. Stud. 2024, 1 (2), 89–103.
25. Attai, K.; Ekpenyong, M.; Amannah, C.; Asuquo, D.; Ajuga, P.; Obot, O.; Johnson, E.; John, A.; Maduka, O.; Akwaowo, C.; Uzoka, F.-M. Enhancing the Interpretability of Malaria and Typhoid Diagnosis with Explainable AI and Large Language Models. Trop. Med. Infect. Dis. 2024, 9 (216). https://doi.org/10.3390/tropicalmed9090216.
26. Su, M.; Guo, J.; Chen, H.; Huang, J. Developing a Machine Learning Prediction Algorithm for Early Differentiation of Urosepsis from Urinary Tract Infection. Clin. Chem. Lab. Med. 2023, 61 (3), 521–529. https://doi.org/10.1515/cclm-2022-1006.
27. Prakash, K. B.; Imambi, S. S.; Ismail, M.; Kumar, T. P.; Pawan, Y. N. Analysis, Prediction and Evaluation of COVID-19 Datasets Using Machine Learning Algorithms. Int. J. 2020, 8 (5), 2199–2204. https://doi.org/10.30534/ijeter/2020/117852020.
28. Kumarakulasinghe, N. B.; Blomberg, T.; Liu, J.; Leao, A. S.; Papapetrou, P. Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS); IEEE, 2020; pp 7–12. https://doi.org/10.1109/CBMS49503.2020.00009.
29. Attai, K.; Akwaowo, C.; Asuquo, D.; Esubok, N. E.; Nelson, U. A.; Dan, E.; Uzoka, F. M. Explainable AI Modelling of Comorbidity in Pregnant Women and Children with Tropical Febrile Conditions. Proc. Int. Conf. Artif. Intell. Appl. 2023, 152–159. https://doi.org/10.59200/ICARTI.2023.022.
30. University of Uyo Teaching Hospital; Mount Royal University. NFRF Project Patient Dataset with Febrile Diseases [Data Set]. Zenodo. https://doi.org/10.5281/zenodo.13756418 (accessed 2024).
31. Yousefi, M.; Rahmani, K.; Rajabi, M.; Reyhani, A.; Moudi, M. Random Forest Classifier for High Entropy Alloys Phase Diagnosis. Afr. Mat. 2024, 35 (3), 57. https://doi.org/10.1007/s13370-024-01198-1.
32. Palimkar, P.; Shaw, R. N.; Ghosh, A. Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In Advanced Computing and Intelligent Technologies; Bianchini, M., Piuri, V., Das, S., Shaw, R. N., Eds.; Springer: Singapore, 2022; Vol. 218, pp 317–327. https://doi.org/10.1007/978-981-16-2164-2_19.
33. Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the Prediction of Student Performance Based on the Machine Learning XGBoost Algorithm. Interact. Learn. Environ. 2021, 31 (6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235.
34. Wu, Y.; Zhang, L.; Bhatti, U. A.; Huang, M. Interpretable Machine Learning for Personalized Medical Recommendations: A LIME-Based Approach. Diagnostics 2023, 13 (16), 2681. https://doi.org/10.3390/diagnostics13162681.

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.