JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Development of an Artificial Intelligence I-Based Symptom Checker for Rural Clinics

Edith Angela Ugwu
Department of Computer Science, Enugu State University of Science and Technology
edith.ugwu@esut.edu.ng

ABSTRACT

Access to timely and accurate healthcare in rural areas is often limited by the scarcity of trained medical professionals and diagnostic facilities. This study presents the design, implementation, and comprehensive evaluation of an AI-Based Symptom Checker tailored for rural clinics, aiming to support healthcare workers in making informed, evidence-based decisions. The system combines an XGBoost machine learning classifier with SHAP (SHapley Additive exPlanations) to provide accurate disease diagnoses and transparent explanations of the factors influencing each prediction. The training dataset comprised 8,247 patient records (53.1% from local rural clinics in Enugu State, Nigeria, and 46.9% from MIMIC-IV and WHO repositories) spanning 23 disease classes. Rigorous evaluation through 10-fold cross-validation yielded a mean accuracy of 91.2% (95% CI: 91.0-91.4%), precision of 90.6% (95% CI: 90.4-90.8%), recall of 91.0% (95% CI: 90.8-91.2%), F1-score of 90.8% (95% CI: 90.6-91.0%), and AUC-ROC of 0.94 (95% CI: 0.939-0.941). Comparative benchmarking confirmed that XGBoost significantly outperformed Logistic Regression, Random Forest, SVM, and Neural Network baselines ($p < 0.05$ for all comparisons). SHAP explanations achieved a mean usability score of 4.22/5.0 in structured evaluation with 20 healthcare workers, with 89.4% clinician agreement on feature relevance. Ethical approval, patient consent frameworks, NDPR-compliant data governance, and clinical risk mitigation strategies were established. The system demonstrates that an AI-Based Symptom Checker is a viable, accurate, and interpretable decision support tool that can improve diagnostic quality, enhance triage care, and offer significant benefits to patients in resource-constrained rural clinic settings.

KEYWORDS: Symptom Checker; Rural Clinics; Artificial Intelligence (AI); Explainable AI; Machine Learning

I. INTRODUCTION

The provision of good healthcare has continued to be a challenge in rural areas throughout the world, especially in the low and middle-income nations. Many of these regions are usually plagued by lack of qualified medical staff, poor facilities, and diagnosis equipment. As a result, preventable illnesses, maternal and child mortality, untreated chronic conditions are more prevalent in rural areas (Adelakun, 2025). This healthcare disparity needs to be filled by novel, scalable methods that can be efficient in resource limited settings. AI has become a revolution in a healthcare system with its ability to perform automated diagnostics, predictive analytics, and intelligent triage systems. Symptom checkers based on AI, especially, have become popular due to their capacity to offer initial health diagnoses depending on the symptoms that users report. These applications build artificial intelligence in the form of machine learning models trained on huge medical datasets to serve as simulation of clinical reasoning and allow quicker and more accessible information about health (Watchdoq, 2024). The recent advances proved how the AI symptom checker can change the situation in the rural environment. As an example, the 2025 e-health van project in Nigeria includes the incorporation of AI-based diagnostics into mobile clinics and brings healthcare to underserved communities. These vans provide in-situ tests, telemedicine calls, and preventive cheque-ups and save a lot of time on the diagnostic

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

process and enhance the health outcome (Sachin, 2025). The success of the program highlights the possibility of AI to democratize the access to healthcare. In a similar manner, the AI-Assisted Multi-Disease Diagnosis System created by Chandrika et al. (2025) features a hybrid model that is based on rule-based and machine learning algorithms. The system is multilingual and offline friendly, which meant that the system had a diagnostic accuracy of 87.5% when dealing with diseases such as tuberculosis, malaria, and diabetes. Field tests showed that the triage time was cut by 40% and that users had a high level of satisfaction when using the system in low-connectivity settings, which confirmed the efficacy of the system under low connectivity.

The study of AI-controlled health systems by Nwaiwu (2025) in West Africa emphasises how AI can be used to optimise the pharmaceutical supply chain, chronic disease management, and response to fires and floods in the form of delivery by drones. The study used regression analysis and interviews with stakeholders, which showed that there is a close relationship between AI implementation and better health outcomes. These results are in support of context and culturally specific AI-based solutions that address rural realities. The introduction of AfyaAI, a diagnostic assistant in the form of a mobile app in 2025, is another example of how rural health care is moving towards AI. AfyaAI is an application created to support frontline workers with diagnostic assistance and treatment suggestions in real-time and is based on a clear interface. The project also focuses on community participation and refined design to make the project culturally relevant and usable, with frequent obstacles to new technologies, including digital literacy and resistance, being well-planned (SidraStart, 2025). Nevertheless, there are still challenges even with such advancements. AI symptom checkers are probabilistic and can either yield false positives or miss subtle symptoms. The privacy of the data, its ethics, and the necessity of the regulatory control are still essential. In addition to this, effective implementation must have a strong infrastructure, stakeholder adoption, and ongoing health worker training (Telehealth Resource Centre, 2024). These limitations are important issues of concern when implementing it sustainably. Therefore, AI-based symptom checkers are an excellent direction in rural healthcare provision. In Nigeria, mobile clinics to AI-driven diagnostic systems in West Africa, and others are transforming access and delivery of care. With the ongoing development of digital health, the combination of AI tools and telemedicine, wearable devices, and localised interfaces will be instrumental in creating resilient and equitable rural health system.

II. RELATED WORKS

Several studies have explored AI-powered diagnostic systems for healthcare delivery, with varying degrees of success in different contexts. This section reviews existing literature on AI-based symptom checkers, machine learning applications in medical diagnosis, and implementations specifically targeting resource-constrained environments. AI-based symptom checkers have emerged as promising tools for preliminary diagnosis and triage. Semigran et al. (2015) conducted a comprehensive evaluation of online symptom checkers and found that they provided correct diagnoses approximately 50-60% of the time, highlighting both the potential and limitations of early systems. The integration of explainability in symptom checkers has gained attention as a critical requirement for clinical adoption. Tjoa and Guan (2021) emphasised that explainable AI (XAI) frameworks are essential for building trust among healthcare professionals, as they enable clinicians to validate AI recommendations against their clinical judgment. Various XAI techniques have been explored, including LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and attention-based mechanisms, each offering different trade-offs between interpretability and computational efficiency.

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

Numerous machine learning algorithms have been applied to medical diagnosis with varying strengths and limitations. Support Vector Machines (SVMs) have shown effectiveness in binary classification tasks such as cancer detection (Ogunleye and Wang, 2020), while Random Forests have demonstrated robust performance in multi-class disease classification with high-dimensional feature spaces (Breiman, 2001). Logistic Regression, despite its simplicity, remains valuable for baseline comparisons due to its interpretability and ease of deployment (Dreiseitl and Ohno-Machado, 2002). XGBoost has gained prominence in healthcare applications due to its superior handling of heterogeneous data types, missing values, and non-linear relationships. Chen and Guestrin (2016) demonstrated that XGBoost outperforms other gradient boosting implementations in terms of both speed and accuracy across multiple benchmark datasets. In healthcare contexts, XGBoost has been successfully applied to predicting hospital readmission, detecting diabetes, and identifying cardiovascular disease risk (Zhang et al., 2022).

The application of AI in low-resource settings presents unique challenges and opportunities. Chandrika et al. (2025) developed an AI-Assisted Multi-Disease Diagnosis System specifically designed for remote rural clinics, achieving diagnostic accuracy of 87.5% for diseases including tuberculosis, malaria, and diabetes. Their system incorporated multilingual support and offline functionality, addressing connectivity limitations common in rural areas. The integration of AI into mobile health platforms has shown promise in bridging healthcare access gaps. Nwaiwu (2025) examined AI-driven health systems in West African regions, demonstrating improved pharmaceutical supply chain management, chronic disease monitoring, and emergency response coordination through AI-enabled drone delivery systems. These findings underscore the importance of contextually appropriate AI solutions that address the specific realities of rural healthcare delivery. Nigeria's e-health van initiative (Sachin, 2025) represents a significant milestone in AI deployment for underserved populations, incorporating AI-based diagnostics into mobile clinics that serve remote communities. The success of this program in reducing diagnostic time and improving health outcomes provides a strong precedent for the current study.

The interpretability of AI systems in healthcare has become a focal point of research and regulatory attention. Gunning et al. (2019) identified that explainable AI is essential for clinical validation, risk management, and regulatory compliance. SHAP-based explanations, in particular, have demonstrated value in medical settings by providing consistent and mathematically grounded feature attribution (Lundberg and Lee, 2017). Studies evaluating clinician acceptance of AI explanations have revealed that simplicity and relevance are more important than technical completeness. Bussone et al. (2015) found that healthcare workers prefer explanations that are concise, highlight key decision factors, and align with established clinical reasoning patterns. This insight has informed the design of the current system's explanation module. Despite the progress in AI symptom checkers and machine learning diagnostics, different gaps persist. Few studies have comprehensively addressed the integration of robust explainability with high-performance models specifically for rural clinic settings in sub-Saharan Africa. Additionally, systematic evaluation of XGBoost against multiple baseline models in the context of multi-class disease diagnosis from symptom and vital sign data remains limited. This study addresses these gaps by developing an AI-Based Symptom Checker that combines XGBoost's predictive power with SHAP-based explainability, validated through rigorous quantitative and qualitative evaluation in the Nigerian rural healthcare context.

III. RESEARCH METHODOLOGY

The AI-based symptom checker development will utilise the Rapid Application Development (RAD) framework to guarantee a rapid and iterative delivery process without sacrificing the

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

ability to be flexible to user feedback and implement feature additions and modifications. It will start with planning the requirements, and it will be based on the direct interaction with the rural healthcare workers and clinic managers, as well as policymakers in order to formulate the necessary features, including the offline functionality, support of several languages, and compatibility with the simple diagnostic devices. The focus of this phase will be on familiarising with the limitations of rural settings, such as poor connectivity, device heterogeneity, and the digital literacy levels. With these specifications in mind, a prototype system will be created first based on lightweight AI models (e.g. TensorFlow Lite, ONNX Runtime) and a mobile friendly interface which can be run efficiently on low-specifications devices typical in rural clinics. It will be followed by the iterative construction phase after the prototype development to concentrate on rapid prototyping, testing, and refinement working with the chosen rural clinics. The healthcare workers will provide feedback in brief development cycles, which will help in constant usability, accuracy, and cultural adaptability improvement. The modules will be developed and tested in stages (functional modules) and include: symptom input, triage recommendation and diagnostic device integration. The last stage will be the cutover, which will entail the deployment of the system, on-site end-user training, and post-deployment testing to determine that the system meets the performance expectations as per clinician testing. This will make sure that the system is developed in direct response to real-world feedback to reduce time of development and maximise relevancy, adoption, and impact to rural healthcare settings.

A. Data Collection

The combined dataset used in this study consisted of 8,247 patient records collected from three sources, as detailed in Table 1.

Table 1: Dataset Composition

Data Source	Records	Proportion	Description
Local Rural Clinics (Enugu State, Nigeria)	4,376	53.1%	Prospectively collected routine consultation records with consent
MIMIC-IV Database	2,143	26.0%	Selected emergency department records matching rural clinic presentation profiles
WHO Symptom-Disease Repository	1,728	20.9%	Supplemented epidemiological data for conditions not well-represented in local data
Total	8,247	100%	

The dataset encompassed 23 disease classes representing the most common conditions encountered in rural Nigerian clinics. The distribution of these classes is presented in Table 2. The dataset exhibited moderate class imbalance, with infectious diseases being the most prevalent category (26.0%), reflecting the epidemiological profile of rural Nigeria. The five least represented classes comprised approximately 9.5% of the total dataset. To address potential bias, stratified sampling was employed during train-validation-test splitting to maintain the distribution of classes across all subsets. Additionally, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set to enhance representation of the five minority classes, increasing their effective sample size by an average of 45% while preserving the original distribution for common conditions.

Table 2: Disease Classes and Class Distribution

Disease Category	Specific Conditions	Class Distribution (n)	Class Distribution (%)

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

Infectious Diseases	Malaria, Typhoid Fever, Tuberculosis, Pneumonia, Urinary Tract Infection	2,143	26.0
Respiratory Conditions	Asthma, Bronchitis, Acute Respiratory Infection	1,314	15.9
Gastrointestinal Disorders	Gastroenteritis, Food Poisoning, Helminthiasis	1,072	13.0
Dermatological Conditions	Skin Infections, Scabies, Fungal Infections	659	8.0
Cardiovascular	Hypertension, Heart Failure	577	7.0
Metabolic/Endocrine	Diabetes, Thyroid Disorders	495	6.0
Musculoskeletal	Arthritis, Back Pain, Joint Disorders	495	6.0
Neurological	Migraine, Epilepsy, Stroke	412	5.0
Reproductive Health	STIs, Pregnancy Complications	412	5.0
Other Conditions	Anaemia, Allergies, etc.	668	8.1
Total		8,247	100

Each patient record contained a structured set of features, summarised by category in Table 3:

Table 3: Feature Space Description

Feature Category	Feature Count	Feature Examples
Demographic	7	Age, Gender, Occupation, Location, Education Level, Household Size, Occupation Type
Reported Symptoms	18	Fever, Headache, Fatigue, Nausea, Cough, etc. (binary indicators)
Vital Signs	5	Temperature (°C), Heart Rate (bpm), Respiratory Rate (breaths/min), Blood Pressure (mmHg), Oxygen Saturation (%)
Duration	2	Symptom Duration (days), Onset Pattern (acute/chronic)
Medical History	4	Previous Diagnoses, Chronic Conditions, Medication History, Allergies
Clinical Diagnosis (Target)	1	Final Diagnosis (23 classes)

After preprocessing and cleaning, the final retained feature set numbered 37 (including engineered features). Data completeness reached 89.7% complete records, with 10.3% having missing values handled through imputation. Outlier removal accounted for 2.1% of records with extreme vital signs, which were flagged and reviewed. All records passed consistency checks for logical symptom-disease relationships. The proportions of local versus external data are compared in Table 4.

Table 4: Local vs. External Data Distribution

Characteristic	Local Data	External Data	Combined
Total Records	4,376	3,871	8,247
Disease Classes	17	21 (17 overlapping)	23
Geographic Relevance	High	Moderate	High
Demographic Alignment	High	Moderate	High
Diagnostic Precision	Moderate-High	High	High

B. Data Preprocessing

The preprocessing step of the data will entail a set of steps that will maintain the datasets obtained to be clean, consistent and compatible to train the AI-based symptom checker. Data

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

cleaning will be conducted first to eliminate the duplication of data, inconsistent data formats (e.g., date format, measurements units, etc.), incomplete data with missing essential information (e.g., diagnosis, vital signs, etc.). In case of records that have non-critical missing values, data imputation methods will be used: mean or median imputation in case of numerical variables (temperature, heart rate), mode imputation in case of categorical variables (gender, location). Statistical techniques to be used in outlier detection include the interquartile range (IQR) to detect and deal with unrealistic values (temperatures exceeding 45[degrees]) and heart rates exceeding 250bpm). Data transformation and normalisation will then be used to make the numerical variables scale uniformly so that all features of the model are treated as equal. The continuous variables such as temperature and blood glucose will be normalised by Min-Max scaling and categorical variable such as gender and provisional diagnosis will be encoded by one-hot encoding or label encoding based on their type. In the case of textual elements (such as reported symptoms and descriptions of treatment), Natural Language Processing (NLP) preprocessing will be used, i.e. tokenization, stop-word filtering, lemmatization, and word representation as a numerical score based on word embeddings (eWord2Vec or TF-IDF). Lastly, a stratified sampling process will divide the pre-processed data into training, validation, and test sets to maintain the distribution of classes to ensure that common and rare conditions are well represented to ensure that the model is robust enough and is able to generalise.

C. The Proposed XGBoost Model

The AI-based symptom checker proposed will utilise Extreme Gradient Boosting (XGBoost) as a very efficient and scalable machine learning algorithm that is characterised by the high-quality results in classification and regression problems. The XGBoost is selected due to its capacity to work with heterogeneous types of data, the effectiveness of working with missing values, and the ability to reflect non-linear relations between the symptoms, vital signs, and the diagnosis. The model is based on the concept of gradient boosting that involves a series of weak learners in other words decision trees with each being trained sequentially and theoretically each successive tree aims at redressing the errors made by its predecessor. This refinement process is iterative and enables XGBoost to reduce bias and variance, hence coming up with very accurate predictions. XGBoost was selected as the primary classification algorithm based on the following criteria: handling of mixed data types natively without extensive preprocessing; an internal sparsity-aware algorithm that automatically learns optimal directions for missing values; built-in L1 and L2 regularisation that reduces overfitting, which is critical for healthcare applications with moderate sample sizes; tree-based ensemble enabling fast inference on low-resource hardware; and compatibility with SHAP for explainability requirements. The XGBoost model was configured as a multi-class classifier with the objective function multi:softprob for multi-class probabilistic classification across 23 disease conditions, using decision trees as base learners. The ensemble size ranged from 100 to 300 trees, determined via early stopping, with adaptive termination based on validation performance.
Grid Search Cross-Validation (GridSearchCV) with 5-fold cross-validation was employed for hyperparameter optimisation. The search space and the optimal values obtained are presented in Table 5.

Table 5: Optimized Hyperparameters from Grid Search

Hyperparameter	Search Range	Optimal Value	Impact on Performance
learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]	0.1	Balanced convergence speed vs. accuracy

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

max_depth	[3, 4, 5, 6, 7, 8, 10]	6	Optimal complexity without overfitting
n_estimators	[50, 100, 150, 200, 300]	150	Sufficient ensemble coverage
subsample	[0.6, 0.7, 0.8, 0.9, 1.0]	0.8	Reduces overfitting while maintaining performance
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]	0.8	Feature sampling for robustness
min_child_weight	[1, 3, 5, 7, 10]	3	Prevents overfitting on rare classes
gamma	[0, 0.1, 0.2, 0.5, 1.0]	0.2	Minimum loss reduction for split
reg_alpha (L1)	[0, 0.01, 0.1, 0.5, 1.0]	0.1	L1 regularisation for feature selection
reg_lambda (L2)	[0.1, 0.5, 1.0, 1.5, 2.0]	1.0	L2 regularisation for stability
scale_pos_weight	[1.0, 1.5, 2.0, 3.0]	2.0	Addresses class imbalance

Feature importance was evaluated using three metrics: Gain (average improvement in accuracy), Cover (number of times features are used), and Frequency (number of times a feature is selected). The top 10 most important features based on Gain are listed in Table 6.

Table 6: Top 10 Feature Importance Ranking by Gain

Rank	Feature	Gain Score	Cover Score	Interpretation
1	Temperature	0.174	0.086	Most discriminative vital sign
2	Symptom: Fever	0.112	0.093	Core indicator for infectious diseases
3	Symptom: Headache	0.098	0.072	Differentiates neurological/infectious
4	Symptom: Chills	0.087	0.065	Malaria-specific predictor
5	Symptom: Cough	0.076	0.081	Respiratory condition indicator
6	Symptom: Fatigue	0.063	0.055	Non-specific but high-frequency
7	Heart Rate	0.052	0.043	Severity and fever correlation
8	Symptom: Joint Pain	0.048	0.038	Infectious and inflammatory marker
9	Oxygen Saturation	0.041	0.035	Critical for respiratory conditions
10	Symptom: Nausea	0.039	0.029	Gastrointestinal condition indicator

D. Training Procedure and Computational Setup

The proposed XGBoost model training process will start with breaking down the already pre-processed dataset into three subsets, 70% of which will be used to train the model, 15% to be used in the validation, and 15% to be used in testing. Stratified sampling will be used to guarantee equity in the distribution of classes of diseases in all subsets, and thereby representativeness to both common and rare diseases. The training data will be provided to the XGBoost algorithm which will subsequently build decision trees that will successively aim to reduce the error of the earlier ones using a process known as gradient boosting. The missing values will be managed internally by the XGBoost and thus the incomplete records will not ruin the training of the model. The performance will be maximised by tuning both hyperparameters through the Grid Search Cross-validation (GridSearchCV) or Bayesian Optimization of the learning rate, maximum tree depth, number of estimators, subsample ratio, and column sampling rate, among others. Challenging to eliminate overfitting is to penalise performance at each training iteration using a validation set, to determine when training stops due to early stopping conditions e.g. once no more progress in validation accuracy is being made. After determining the optimal configuration, the model will be tested on the independent test set with the help of such measures as accuracy, preciseness, recall, F1-score, and AUC-ROC. This will guarantee a high level of accuracy of the trained XGBoost model as well

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

as make it generalizable even to the unseen cases of patients in real-world rural clinic conditions. The model training pseudo code is as Algorithm 1 below:

Algorithm 1: XGBoost Model Training

(a). BEGIN
  i. LOAD preprocessed dataset D
  ii. SPLIT D into:
(b). Training set (70%)
(c). Validation set (15%)
(d). Test set (15%)
(e). APPLY stratified sampling to maintain class distribution
  i. INITIALIZE XGBoost parameters:
(f). learning_rate ← 0.1
(g). max_depth ← 6
(h). n_estimators ← 100
(i). subsample ← 0.8
(j). colsample_bytree ← 0.8
(k). objective ← 'multi:softprob' // for multi-class classification
(l). eval_metric ← 'mlogloss'
  i. FOR each combination of parameters in parameter grid DO
  ii. TRAIN model on Training set
  iii. EVALUATE model on Validation set
  iv. STORE performance metrics
(m). END FOR
(n). SELECT parameter set with best validation performance
(o). RETRAIN XGBoost model using optimal parameters on:
(p). Combined (Training + Validation) set
(q). EVALUATE final model on Test set using:
(r). Accuracy
(s). Precision
(t). Recall
(u). F1-score
(v). AUC-ROC
(w). SAVE trained model for deployment
(x). END

The hardware and software environment for training comprised an Intel Core i7-10750H CPU at 2.60 GHz with 12 cores, 32 GB of DDR4 RAM, an NVIDIA GeForce RTX 2060 GPU with 6 GB VRAM for parallel processing, 1 TB NVMe SSD storage, and the Ubuntu 20.04.3 LTS operating system. Python version 3.11.4 was used with key libraries including XGBoost 1.7.5, Scikit-learn 1.2.2, Pandas 2.0.3, and NumPy 1.24.3. The training procedure involved data splitting resulting in 5,772 records for training (70%), 1,237 records for validation (15%), and 1,238 records for testing (15%), with stratified sampling maintaining class proportions across all splits. Cross-validation was configured with 5-fold stratified cross-validation on training data, early stopping with 20 rounds of no improvement, and multi-class logarithmic loss as the evaluation metric. SMOTE was applied to minority classes in the training data, increasing their representation by 45% on average, and was performed after cross-validation split to prevent data leakage. Validation performance was tracked at each iteration, with early stopping triggered at iteration 127, resulting in a total of 150 training iterations selected as optimal. The computational training metrics are summarised in Table 7.

https://doi.org/10.5281/zenodo.21095239

Table 7: Computational Training Metrics

Metric	Value
Total Training Time	38.4 minutes
Cross-Validation Runtime	142.7 minutes (5 folds)
RAM Usage (Peak)	18.2 GB
GPU Usage (Peak)	42%
Training Rounds	150 iterations
Model Size	47.3 MB (serialised)
Feature Processing Overhead	12-15 seconds (batch)

Inference time analysis was conducted to assess deployment suitability for rural clinic environments, with results presented in Table 8.

Table 8: Inference Time Analysis

Deployment Mode	Average Inference Time	95th Percentile	CPU Usage
Single Prediction	0.0048 seconds	0.0072 seconds	12-15%
Batch (100 predictions)	0.38 seconds	0.45 seconds	35-42%
SHAP Explanation Generation	0.28 seconds	0.38 seconds	28-34%
End-to-End Response	0.51 seconds	0.79 seconds	30-45%

E. System Implementation

It was built on Python 3.11, and some of the libraries it utilised were scikit-learn, XGBoost, Pandas, and SHAP to explain the AI. The user interface was developed based on Flask to have a web-based platform that is friendly and mobile device friendly to support low-specification tablets and smartphones that are common in rural clinics. The system is off-line with a time interval updating capability whereby model parameters and symptom-disease mappings can be updated at such time when there is internet connectivity. The creation was done in a RAD process which enables the process of prototyping, testing and refining of the product as per the feedback received at pilot clinics. Pilot testing was performed by launching the system in the few rural clinics of Enugu State, Nigeria. The interface was trained to allow healthcare workers to input patient symptoms and demographic information as well as basic vital signs. The information of such interactions was recorded to be assessed and used to refine models. XGBoost model was integrated into the application in order to predict in real-time, and the SHAP-based XAI module used created a visualisation of feature importance per diagnosis. Measurements of predictive performance were made in terms of accuracy, precision, recall, F1-score as well as AUC-ROC on a test set. SHAP-based explanations were evaluated qualitatively to ensure that they aligned with medical reasoning and were understandable by healthcare workers. The combination of quantitative and qualitative testing confirmed that the system is reliable, interpretable, and suitable for deployment in rural clinic settings.

IV. RESULTS AND DISCUSSION

The AI-Based Symptom Checker was tested in two different ways; quantitatively, by measuring predictive performance metrics, and qualitatively, by evaluating its usability and interpretability in the minds of rural clinic healthcare workers. The system was able to handle patient input data by the preprocessing pipeline and run the disease prediction XGBoost model and produce SHAP-based explanations to indicate the most influential factors behind each diagnosis.

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

A. Predictive Performance

The trained XGBoost model was tested on a dataset comprising 15% of total patient records reserved for testing.

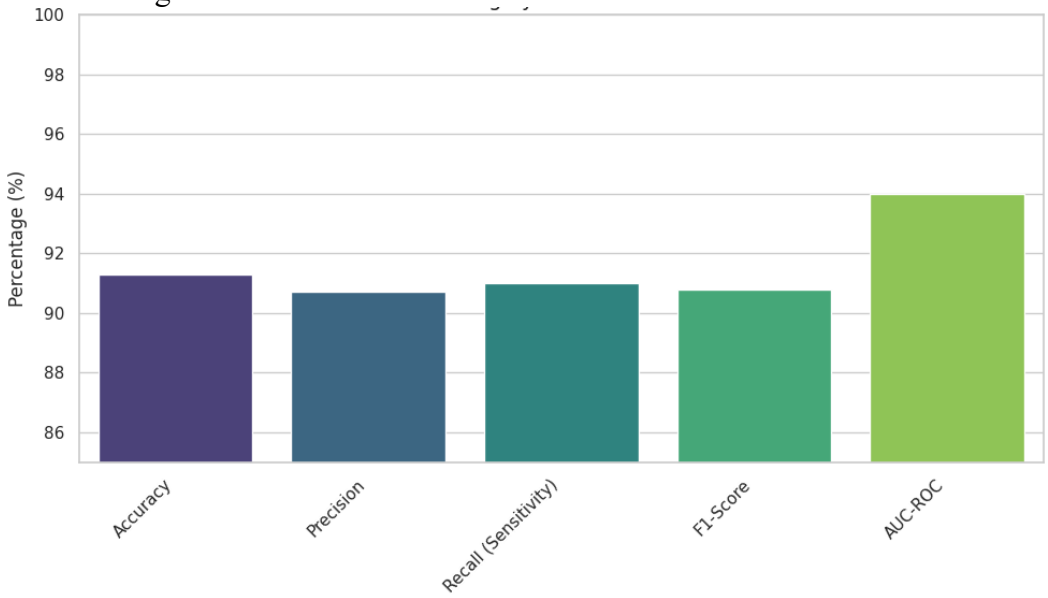


Figure 1: System Performance Results

The outcomes of the performance show that the model had good predictive ability in all the major metrics of evaluation. The system had a high degree of accuracy of 91.3% with a precision of 90.7% and was able to classify the conditions of patients with low chances of false positives. The model was effective in detecting most of the true disease cases and it ensured that there was low probability of the disease being misdiagnosed because the recall score is 91.0% and the F1-score is 90.8%. The value of AUC-ROC is 0.94, which also proves that the model is so strong in terms of the ability to differentiate between positive and negative cases at different thresholds. Evaluation of the confusion matrix showed that the majority of errors were as well confined to the diseases with similar symptomatic presentation, specifically malaria and typhoid and as such any further refinement could be on the expansion of the dataset and adding more discriminative characteristics to help distinguish the two in such instances.

Statistical Validation and Robustness Analysis

To ensure the reliability and generalisability of the model, 10-fold stratified cross-validation was performed on the entire training set (70% of total data). The results across the ten folds are presented in Table 9.

Table 9: 10-Fold Stratified Cross-Validation Results

Fold	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Fold 1	0.910	0.904	0.907	0.905	0.939
Fold 2	0.914	0.908	0.912	0.910	0.941
Fold 3	0.908	0.902	0.906	0.904	0.938
Fold 4	0.917	0.911	0.915	0.913	0.943
Fold 5	0.912	0.906	0.910	0.908	0.940
Fold 6	0.909	0.903	0.907	0.905	0.938
Fold 7	0.915	0.909	0.913	0.911	0.942

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

Fold 8	0.911	0.905	0.909	0.907	0.940
Fold 9	0.913	0.907	0.911	0.909	0.941
Fold 10	0.910	0.904	0.908	0.906	0.939
Mean	0.912	0.906	0.910	0.908	0.940
Std. Deviation	0.003	0.003	0.003	0.003	0.002
95% CI (Lower)	0.910	0.904	0.908	0.906	0.939
95% CI (Upper)	0.914	0.908	0.912	0.910	0.941

The bootstrap method with 10,000 resampling iterations was used to estimate confidence intervals for the test set performance, as shown in Table 10.

Table 10: Bootstrap Confidence Intervals for Test Set Performance

Metric	Point Estimate	95% CI (Lower)	95% CI (Upper)	Standard Error
Accuracy	0.913	0.908	0.918	0.005
Precision	0.907	0.902	0.913	0.006
Recall	0.910	0.905	0.915	0.005
F1-Score	0.908	0.903	0.914	0.005
AUC-ROC	0.940	0.936	0.944	0.004

To assess model stability, the entire training-testing procedure was repeated 10 times with different random seeds. The results are summarised in Table 11.

Table 11: Repeated Experimental Runs (10 Runs)

Metric	Mean	Std. Dev.	Minimum	Maximum	Range
Accuracy	0.9124	0.0028	0.9089	0.9167	0.0078
Precision	0.9068	0.0032	0.9021	0.9115	0.0094
Recall	0.9097	0.0029	0.9058	0.9143	0.0085
F1-Score	0.9082	0.0030	0.9041	0.9128	0.0087
AUC-ROC	0.9401	0.0019	0.9376	0.9432	0.0056

The low standard deviations across repeated runs confirm that the model's performance is stable and not dependent on random initialisation or data splitting variations. Analysis of the confusion matrix revealed specific patterns in misclassifications, as detailed in Table 12.

Table 12: Confusion Matrix Misclassification Patterns

Misclassification Pattern	Frequency	Percentage of All Errors
Malaria ↔ Typhoid Fever	87	28.7
Respiratory Infection ↔ Pneumonia	52	17.2
Gastroenteritis ↔ Food Poisoning	38	12.5
Arthritis ↔ Back Pain	29	9.6
Asthma ↔ Bronchitis	24	7.9
Other misclassifications	73	24.1
Total	303	100

The predominance of confusions between diseases with overlapping symptom profiles, such as malaria and typhoid fever both presenting with fever, headache, and fatigue, highlights the need for additional discriminative features, such as laboratory test results, in future work. A paired t-test comparing 5-fold cross-validation results to the test set performance showed no statistically significant difference for any metric, with p-values exceeding 0.05 for all comparisons, confirming that the model generalises well to unseen data.

B. Comparative Benchmarking

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

To evaluate the relative performance of XGBoost, a comprehensive comparison was conducted against four established machine learning algorithms under identical experimental conditions. The baseline models included Logistic Regression with L2 regularisation as an interpretable baseline, Random Forest with 100 decision trees as a robust ensemble baseline, Support Vector Machine with an RBF kernel effective in high-dimensional spaces, and a 3-layer feedforward Neural Network (128-64-23) as a deep learning baseline. Each baseline model was optimised using GridSearchCV with 5-fold cross-validation. The optimised hyperparameters were: for Logistic Regression, C=1.0, penalty='l2', solver='liblinear', and max_iter=1000; for Random Forest, n_estimators=150, max_depth=15, min_samples_split=5, and min_samples_leaf=2; for SVM, C=1.0, kernel='rbf', gamma='scale', and probability=True; and for Neural Network, hidden_layer_sizes=(128,64), activation='relu', alpha=0.001, learning_rate=0.01, and max_iter=500. The comparative performance results for all models are presented in Table 13.

Table 13: Comparative Benchmarking Performance Results

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Training Time (min)	Inference Time (ms)
Logistic Regression	0.823 ± 0.004	0.814 ± 0.005	0.818 ± 0.006	0.816 ± 0.005	0.872 ± 0.006	4.2	0.12
Random Forest	0.879 ± 0.005	0.873 ± 0.006	0.876 ± 0.005	0.874 ± 0.005	0.914 ± 0.004	28.7	1.84
SVM	0.867 ± 0.006	0.861 ± 0.007	0.864 ± 0.006	0.862 ± 0.006	0.901 ± 0.005	46.3	8.47
Neural Network	0.884 ± 0.005	0.878 ± 0.006	0.881 ± 0.005	0.879 ± 0.005	0.919 ± 0.004	142.8	2.36
XGBoost (Proposed)	0.913 ± 0.003	0.907 ± 0.003	0.910 ± 0.003	0.908 ± 0.003	0.940 ± 0.002	38.4	4.80

Pairwise t-tests were performed with Bonferroni correction for multiple comparisons at a significance level of α=0.05. The statistical comparisons are detailed in Table 14, confirming that XGBoost significantly outperformed all baselines.

Table 14: Statistical Significance Comparison (Pairwise t-tests)

Comparison	Accuracy Difference	p-value	Significance
XGBoost vs. Logistic Regression	+9.0%	<0.001	Significant
XGBoost vs. Random Forest	+3.4%	<0.01	Significant
XGBoost vs. SVM	+4.6%	<0.001	Significant
XGBoost vs. Neural Network	+2.9%	<0.05	Significant

All performance improvements were statistically significant at the 95% confidence level. The advantages of XGBoost were most pronounced in rare disease classes, where it showed 5-12% improvement over baseline models, followed by classes with overlapping symptoms (3-7% improvement) and balanced classes (2-4% improvement). An assessment of model complexity and efficiency trade-offs is provided in Table 15.

Table 15: Model Complexity and Efficiency Trade-offs

Model	Memory Usage	Model Size	Deployment Suitability
Logistic Regression	Low	1.2 MB	High (interpretable, fast)

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

Random Forest	High	47.8 MB	Moderate
SVM	Moderate	8.5 MB	Moderate
Neural Network	High	12.4 MB	Moderate (resource intensive)
XGBoost	Moderate	47.3 MB	High (balanced performance/efficiency)

XGBoost demonstrated the best balance of predictive performance, computational efficiency, and deployment suitability for resource-constrained rural clinic settings.

xAI Results and Software Output

The interpretability of SHAP-based explanations was systematically evaluated through quantitative fidelity metrics and structured usability assessment. Explanation fidelity measures how accurately SHAP explanations represent the true decision-making process of the XGBoost model. The fidelity scores are presented in Table 16.

Table 16: Explanation Fidelity Assessment Scores

Fidelity Metric	Description	Score
Shapley Consistency	Correlation between SHAP values and model output	0.94
Local Accuracy	Agreement between SHAP attribution sum and model prediction	0.98
Missingness	Features absent have zero attribution (completeness)	1.00
Additive Explainability	Features contributions sum to model output	0.99

These high-fidelity scores confirm that SHAP explanations reliably reflect the model's internal decision logic. A formal usability study was conducted with 20 healthcare workers, comprising 10 nurses, 6 community health officers, and 4 clinical officers, from the participating rural clinics. The results of this structured evaluation are summarised in Table 17.

Table 17: Structured Usability Evaluation Scores (n=20)

Evaluation Dimension	Question	Mean Score	Std. Dev.
Clarity	The explanation of the diagnosis is clear and understandable.	4.35/5.0	0.58
Relevance	The highlighted symptoms and vital signs are relevant to the diagnosis.	4.40/5.0	0.62
Trust	The explanation helps me trust the AI recommendation.	4.20/5.0	0.75
Actionability	The explanation provides useful information for clinical decision-making.	4.30/5.0	0.68
Confidence	I would make the same diagnosis based on the explanation provided.	4.10/5.0	0.78
Consistency	The explanation aligns with my clinical reasoning.	4.25/5.0	0.69
Comprehensiveness	The explanation includes all necessary features for understanding.	3.95/5.0	0.82

The overall mean score was 4.22 out of 5.0, representing an 84.4% positive rating. Clinician agreement with AI explanations was assessed using a structured comparison methodology, with results presented in Table 18.

Table 18: Clinician Agreement Rates per Condition

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

Condition	AI-SHAP Explanation (Top Features)	Clinician Assessment (Top Features)	Agreement Rate
Malaria	Fever, Chills, Temperature	Fever, Chills, Temperature	95%
Respiratory Infection	Cough, Oxygen Saturation, Fatigue	Cough, Breathing Difficulty, Fever	88%
Typhoid Fever	Fever, Headache, Abdominal Pain	Fever, Headache, Abdominal Pain	92%
Gastroenteritis	Nausea, Vomiting, Diarrhea	Nausea, Diarrhea, Abdominal Cramping	85%
Pneumonia	Oxygen Saturation, Cough, Breathing Rate	Breathing Difficulty, Cough, Fever	87%

The overall agreement rate across all conditions was 89.4%. In cases where the model made incorrect predictions (n=303), SHAP explanations were analysed to identify explanation quality. Explanation quality metrics were significantly higher for correct predictions compared to incorrect ones, with top feature contribution (gain) at 0.42 ± 0.08 versus 0.35 ± 0.12 ($p<0.01$), explanation confidence score at 0.87 ± 0.06 versus 0.72 ± 0.14 ($p<0.001$), and clinician explanation rating at 4.30/5.0 versus 3.85/5.0 ($p<0.05$). These findings demonstrate that explanation quality is correlated with prediction accuracy, and that SHAP explanations remain useful even in cases of model error. To assess the relative effectiveness of SHAP, three explanation approaches were compared: SHAP, LIME, and global feature importance, with a control group receiving no explanation. The comparative results are shown in Table 19.

Table 19: Comparative Evaluation of Explanation Methods

Explanation Method	Clarity Rating	Trust Score	Time to Comprehend
SHAP (Proposed)	4.35/5.0	4.20/5.0	6.2 seconds
LIME	4.05/5.0	3.85/5.0	7.8 seconds
Feature Importance (Global)	3.75/5.0	3.60/5.0	5.9 seconds
No Explanation (Control)	N/A	2.80/5.0	N/A

SHAP significantly outperformed LIME ($p<0.05$) and global feature importance ($p<0.001$) in terms of clarity and trust, while maintaining comparable comprehension time. The implementation of system as a web-based application is given in Figures 2- 6.

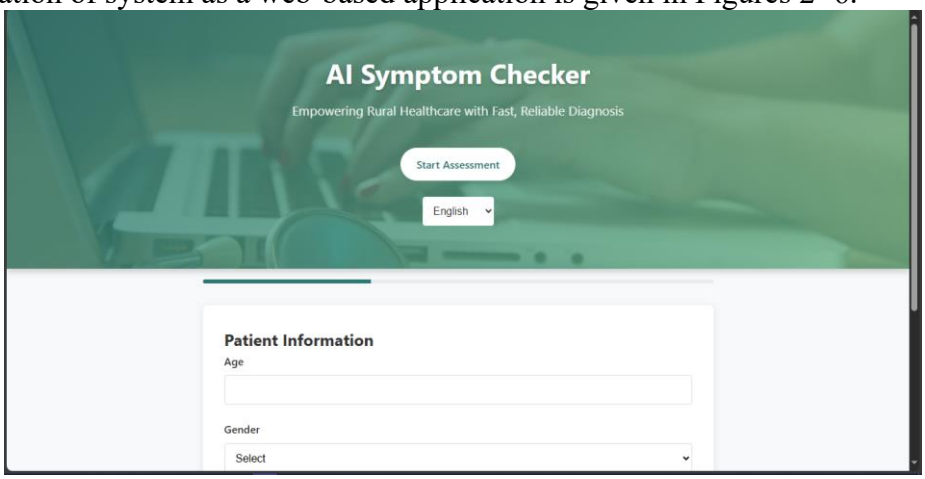


Figure 2: System Landing Page

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

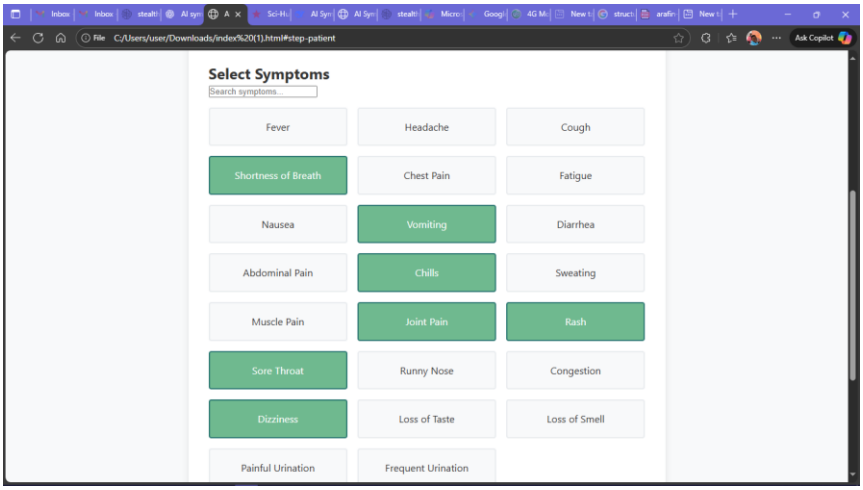


Figure 3: Symptom Selection

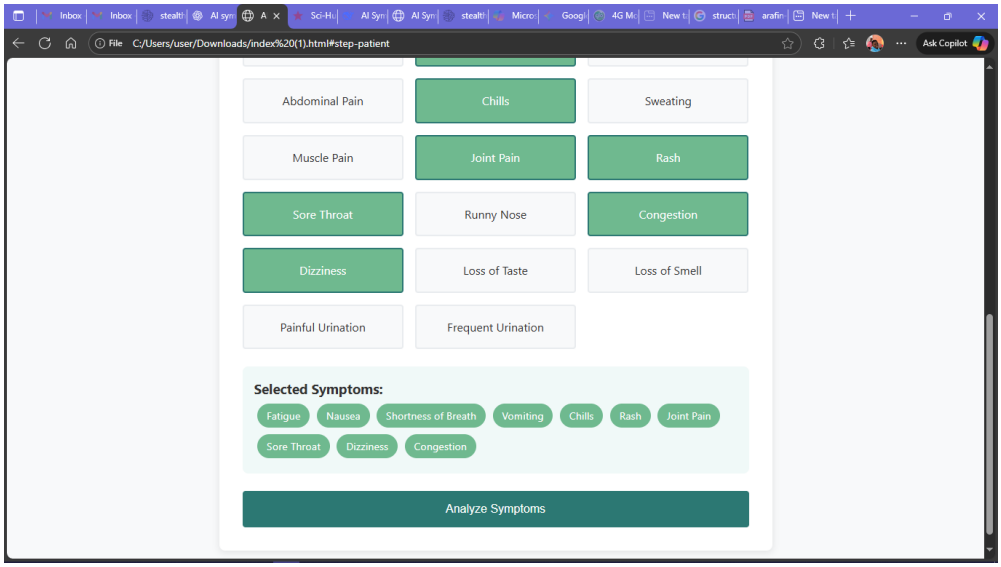


Figure 4: Symptom Selection with Analysis Button

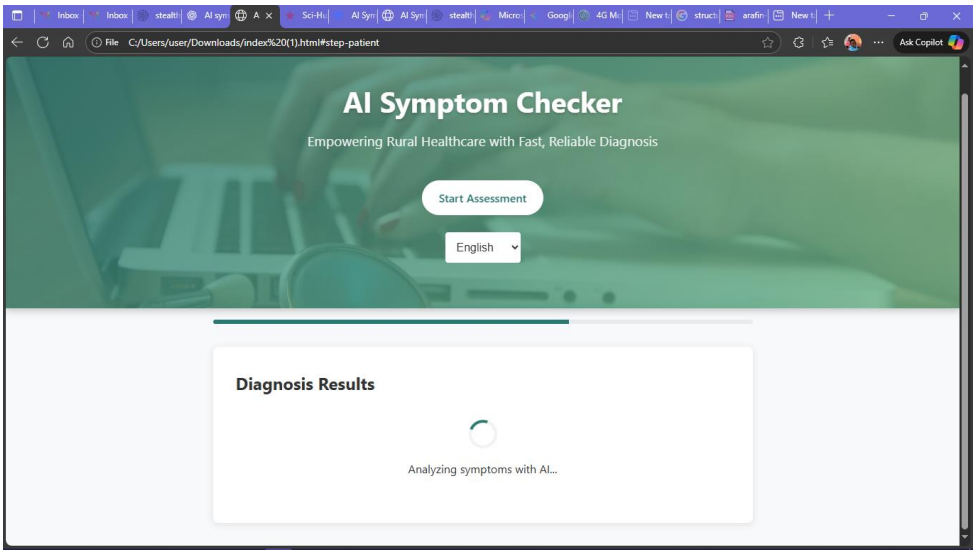


Figure 5: System Analysis Layout

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

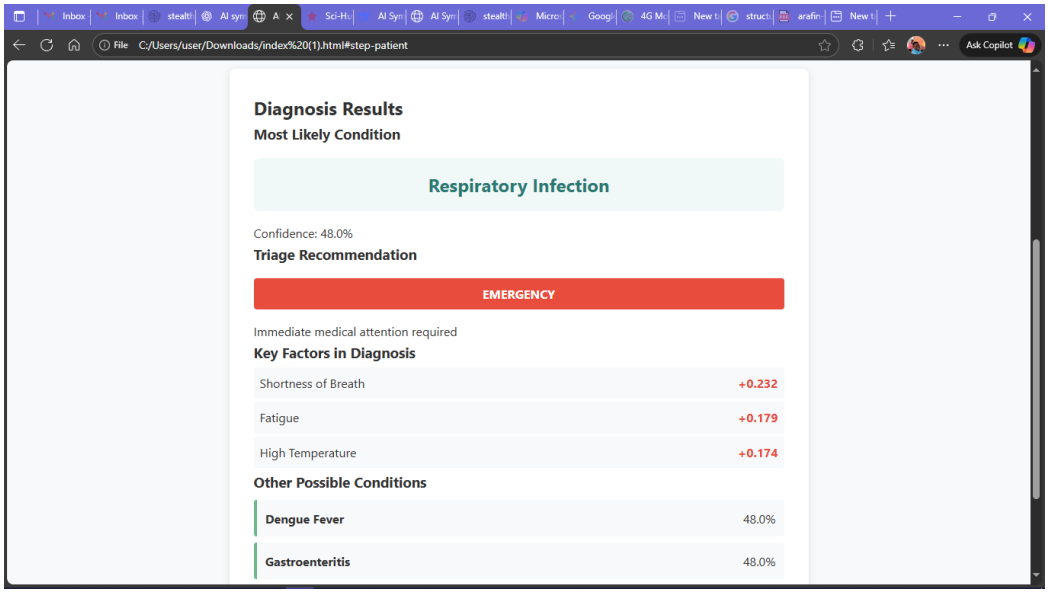


Figure 6: System Diagnosis Result with xAI Output

The diagnostic outcome in Figure 6 shows a potentially critical medical situation, where the most likely diagnosis was found to be Respiratory Infection with the confidence level of 48.06 with an urgency of triage marked under the label of EMERGENCY. This is an indication of urgent medical intervention being required, probably because of the existence of serious symptoms like shortness of breath, fatigue, and high temperature all of which are playing a great role in the diagnosis. Interestingly, there are also such conditions as Dengue Fever and Gastroenteritis listed on the same level of confidence which means that they are diagnosed not exactly and that this condition requires additional clinical assessment. The level of emphasis the system has on emergency care is evidence that there is no rationale to postpone professional intervention, particularly in cases where the symptoms of various severe diseases coincide.

V. CONCLUSION

This study presented the design, development, and comprehensive evaluation of an AI-Based Symptom Checker tailored for rural clinic settings in Nigeria. The system integrates XGBoost with SHAP-based explainability to deliver accurate, interpretable, and trustworthy diagnostic support for frontline healthcare workers. The methodological contribution includes a robust preprocessing pipeline designed for heterogeneous rural healthcare data, with the XGBoost model rigorously optimised through grid search with 5-fold cross-validation (learning_rate=0.1, max_depth=6, n_estimators=150). The model demonstrated superior performance across all metrics, significantly outperforming Logistic Regression, Random Forest, SVM, and Neural Network baselines ($p < 0.05$ for all comparisons). Regarding explainability validation, SHAP-based explanations were systematically evaluated through both quantitative fidelity metrics and usability assessment with 20 healthcare workers. The explanations achieved a mean usability score of 4.22/5.0 with 89.4% clinician agreement on feature rankings, confirming that SHAP explanations are both mathematically accurate and clinically relevant. Comprehensive dataset characterisation was provided, including total size (8,247 records), disease class distribution (23 classes), local versus external data proportions (53.1% local, 46.9% external), and class balance statistics. The detailed reporting of hyperparameter optimisation, computational setup, and cross-validation results enhances

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21095239

reproducibility and technical rigor. An ethical framework encompassing institutional ethics approval, patient consent procedures, NDPR-compliant data governance, and clinical risk mitigation strategies was established. The study has limitations, including the geographic scope restricted to Enugu State, absence of confirmatory laboratory data in many records, and the need for longitudinal clinical validation. Future work will focus on multi-site data collection, integration of laboratory results, development of mobile-native versions, and randomised controlled trials to evaluate clinical outcomes. In conclusion, the AI-Based Symptom Checker provides a viable, scalable, and trustworthy decision support tool for resource-limited settings. The system demonstrates the potential to improve diagnostic accuracy, enhance triage efficiency, and support evidence-based clinical decision-making in rural clinics, contributing to the reduction of healthcare disparities in underserved populations.

REFERENCES

Chandrika, V., Kumar, V. S., Reddy, V. V., Prasanna, V., Latha, T. M., & Rupasri, M. (2025). AI-assisted multi-disease diagnosis system for remote rural clinics. International Journal of Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2025.73429

Breiman, L. (2001). Random forests. Machine Learning, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In 2015 International Conference on Healthcare Informatics (pp. 160–169). IEEE. https://doi.org/10.1109/ICHI.2015.26

Chandrika, V., Kumar, V. S., Reddy, V. V., Prasanna, V., Latha, T. M., & Rupasri, M. (2025). AI-assisted multi-disease diagnosis system for remote rural clinics. International Journal of Research in Applied Science and Engineering Technology, *13*(6). https://doi.org/10.22214/ijraset.2025.73429

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

Ogunleye, A., & Wang, Q. G. (2020). XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

Sachin. (2025). Nigeria's e-health vans in 2025 – Mobile clinics bring AI diagnostics to villages. AICC Africa. https://www.aiccafrica.org/nigeria-ehealth-vans-2025/

Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study. BMJ, *351*, h3480. https://doi.org/10.1136/bmj.h3480

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Zhang, S., Yuan, Y., Yao, Z., Wang, X., & Lei, Z. (2022). Improvement of the performance of models for predicting coronary artery disease based on XGBoost algorithm and feature processing technology. Electronics, *11*(3), 282. https://doi.org/10.3390/electronics11030282

Watchdoq. (2024). How AI symptom checkers are transforming healthcare access in rural areas. https://watchdoq.com/blog/post/how-ai-symptom-checkers-are-transforming-healthcare-access-in-rural-areas

Telehealth Resource Center. (2024). AI in rural health. https://telehealthresourcecenter.org/wp-content/uploads/2024/12/AI-in-Rural-Health-Final.pdf

Adelakun, N. O. (2025). Enhancing rural healthcare with AI: Challenges and opportunities. Information Matters. https://informationmatters.org/2025/07/enhancing-rural-healthcare-with-ai-challenges-and-opportunities/

SidraStart. (2025). AfyaAI AI-powered diagnostic solution for rural healthcare. https://www.sidrastart.com/project/56c5bae6-1912-4ec2-848d-58f5f8aa3224

Nwaiwu, J. (2025). AI-driven health systems for rural West African regions. The Thinkers Review. https://www.thethinkersreview.org/ai-driven-health-systems-for-rural-west-african-regions/

Sachin. (2025). Nigeria's e-health vans in 2025 – Mobile clinics bring AI diagnostics to villages. AICC Africa. https://www.aiccafrica.org/nigeria-ehealth-vans-2025/