Rural Area Health Care System: Disease Prediction with the Symptoms

Uppara Sai Srinivas[1], M. Kameswara Rao[2], Vari Anitha[3] and Venigalla Vaishnavi[4]
{2100050033@kluniversity.in[1], kamesh.manchiraju@kluniversity.in[2], 2100050028@kluniversity.in[3], 2100050056@kluniversity.in[4]}

Department of Electronics and Computer Science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur Dist- 522302, AndhraPradesh, India[1, 2, 3, 4]

Abstract. Access to timely and accurate health services in rural areas is often limited due to the scarcity of medical professionals, lack of diagnostic tools and low health awareness. To address these challenges, this project presents a rural health system: web-based platform-based disease forecast that takes advantage of machine learning to help early detection of common diseases using symptom-based inputs. Currently, the system predicts five primary conditions: hypertension, diabetes, cold types, asthma and gastroenteritis, using appropriate algorithms such as random forest and logistics regression. Each model of disease is trained in relevant medical symptoms, allowing rapid and informative predictions without the need for medical experts. The platform is designed with simplicity and offline functionality in mind, the platform allows rural health professionals to enter the patient's symptoms and receive instant predictive feedback, reducing diagnostic time and workload. Additional planned resources include healthcare professional login to manage patient history, symptom tracking, medical suggestions and precautions, local accessibility translation and AI - powered chatbot for instant consultation resolution. These enhancements aim to make the system not only a diagnostic tool, but also a complete rural health management assistant, finally improving accessibility, awareness and results of medical care in needy regions.

Keywords: Rural health care system, Web-based platform, Offline support, Machine learning, Disease Forecast / Prediction, Symptom based input, Diagnostic tool.

1 Introduction

Access to healthcare in rural and remote regions remains a significant challenge, especially in developing countries where medical infrastructure is limited. These areas generally lack sufficient healthcare professionals, diagnostic tools, and health awareness programs, leading to delays in the identification and treatment of common diseases. The consequences are severe conditions that are easily manageable in urban environments may escalate to life-threatening stages in rural communities due to late detection and inadequate care. Addressing this critical problem requires economic, scalable and easy - to - use technological inventions that can work even in resource limited environments.

The Rural Health System: Symptom diseases forecast is developed as a web application based on machine learning to support early diagnosis and health awareness in needy areas. This system supports the prediction of the common disease according to the symptoms submitted by the user by a comfortable and user aimed web interface and So on. Presently, five categories of diseases are supported in the system: hypertension, diabetes, and cold -type classification, asthma and gastroenteritis. Each disease model is then trained on medical data sets with the help of supervised learning algorithms, and the resources are selected very carefully to keep the

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357926

symptoms relevant and the model accurate. The backend of the application is written in Python and Flask framework, as well as regular web technology for broader access and user adaption.

A notable aspect of this system is its ability to work offline so that rural health clinics can use it without needing reliable internet access. Users can input symptoms through simplified forms, and the system gives real-time predictions to help health professionals or community volunteers conduct initial assessments. Furthermore, the system incorporates an education module to raise public awareness of typical manifestations, especially in the case of febrile disease.

As it merges a machine learning, offline capabilities, and an intuitive interface, the proposed system presents an application capable of making an impact in the provision of rural healthcare services. Aids in prevention and early detection of diseases; encourage healthy lifestyle. Uncovering such project is the first step, next step should be working on enhancing it by means of mobile application creation, Multilanguage support, usage of other diseases or the bond possibility with electronic health records. Lastly, the system seeks to decrease the service burden on the healthcare facilities, improve the intelligence level of rural-urban medical institutions, and use wise and inexpensive technology.

2 Literature Survey

Widespread exploration of medical prediction model, methods, and technology framework for the development of an efficient and feasible illness prediction system among rural context requires exploration of existing models/approaches and the respective technology framework,similar challenges have been addressed in the recent past.

X. Chen, Y. Liu, W. Zhao [1] in this paper emphasized the utilization of random forest and Support Vector Machine (SVM) algorithms to group pulmonary diseases, such as diabetes and hypertension, based on symptomatic and clinical information of patients from rural areas. It highlighted the random forest's capability to deal with multiple features and noisy information. From this, the use of random forest was considered in the diabetes forecast modules and blood pressure of the project, especially due to its robustness, accuracy and interpretability, which are crucial in low -resource rural environments.

S. Kumar and R. Rajasekaran [2], their research focuses on the use of logistic regression in prediction of gastrointestinal diseases, highlighting its low computational costs and rapid response time. These symptoms align with the project's project requirement for the offline, fast and efficient systems. Logistic regression was considered in the gastroenteritis module, which provides practical balance between model complexity and reliability.

A. Reddy, M. S. Reddy, and T. R. Rao [3], the authors applied Random Forest to differentiate respiratory diseases such as asthma and cold, in accordance with symptom patterns. The results of this led to Random Forest being utilized within the asthma and cold type modules of the project, as it is highly resistant to noise and can distinguish imbalanced feature importance.

A. Prakash and K. Rani [4], this letter introduced an integrated platform for symptoms-based multi-seed diagnosis with emphasis on purpose and access. From this, the design of a simple, general user interface was considered in the project, allowing users to reach many disease prediction devices from a single dashboard.

N. Sharma, P. Verma, and S. Saxena [5], explored the use of decision trees and ensemble methods for classification based on symptoms. The emphasis they placed on ensemble methods like Random Forest buttressed our choice of using them to manage multiple disease predictions for rural applications.

P. Singh and J. Kaur [6], this research used random forests to detect early-stage diabetes using a limited number of input resources, making it suitable for low-data environments. This directly impacted on our choice to use Random Forest for our diabetes module.

R. Gupta, M. Patel, and N. Sharma [7], this research developed the determination of cold types based on analyzing symptom severity employing Random Forest and Gradient Boosting. We built upon their process and employed symptom-weighted reasoning with Random Forest to further distinguish between severe and mild cold infection.

S. Patil and S. Kulkarni [8], this paper emphasized the usage of CSV datasets for offline model education and facts dealing with suitable for rural clinics with confined internet access. From this, CSV-based information processing become adopted inside the device, permitting the models to function without cloud infrastructure.

Y. Zhang, L. Wu, and T. Lin [9], the researchers developed a Flask-based web app integrating ML fashions for health prediction. This supported the usage of Flask in the proposed system's backend, enabling a responsive, lightweight interface suitable for low-aid clinics.

S. Rao and D. Kumar [10], in this paper they studied integrating Random Forest and Logistic Regression to construct scalable diagnostic systems. Their work impacted our decision to adopt both models Random Forest and Logistic Regression depending on the disease type and dataset characteristics.

A. Ahmed, H. N. Khan, and S. Farooq [11], this research studied ensemble learning to predict five prevalent diseases using different classifiers. From this, we modified our system architecture to accommodate modular, disease-specific models, enhancing both accuracy and maintenance.

V. Mishra and A. Das [12], the authors designed a scalable rural health system by using open datasets with SVM and Random Forest as disease prediction tools. It encouraged us to structure our system such that local symptoms are fetched and models are trained independently for each disease, thus making our system adaptable and forward-looking.

A. Nair, P. Menon, and R. Thomas [13], the authors created a fever diagnostic assistant focused on educating users rather than predicting illness directly. From this, an informative module was considered in the fever section to raise awareness rather than provide diagnosis.

D. Basak and A. Ghosh [14], this research emphasized the need of local-language interfaces for health tech in rural communities. From this, plans to expand the project with multilingual support were considered to enhance usability for non-English-speaking users in rural areas.

3 Methodology

The Rural Area Health Care System: Disease Prediction with Symptoms follows a well-established methodology encompassing data collection, preprocessing, model selection, system development, and deployment. An app that is reliable, lightweight, and usable offline that diagnoses the diseases after the user has given input of his/her symptoms. It affords fast, real-time prediction. The following sub-section discusses the design and implementation of system.

Clinically, the first related datasets were obtained from publicly available sources, in which the relevant data sets were about each target disease. For preliminary input features quality control and consistency, the preprocessing of data was performed. It entailed handling missing values, categorizing data, normalizing of numerical fields where necessary and selecting the most useful attributes for the prediction. Both of the conditions have a distinct range of symptoms, which have been chosen by their medical relevance and dataset coverage.

Predicted five diseases: hypertension, diabetes, cold-type classification, asthma, and gastroenteritis. A disease-specific model was trained for each disease. For hypertension, diabetes, cold, and asthma prediction, Random Forest classifiers were adopted due to their strong generalization and superior performance of highly-dimensional medical data. Logistic Regression was selected as the approach for predicting gastroenteritis due to its interpretive nature and good performance for binary classification task with few features.

3.1 Architecture

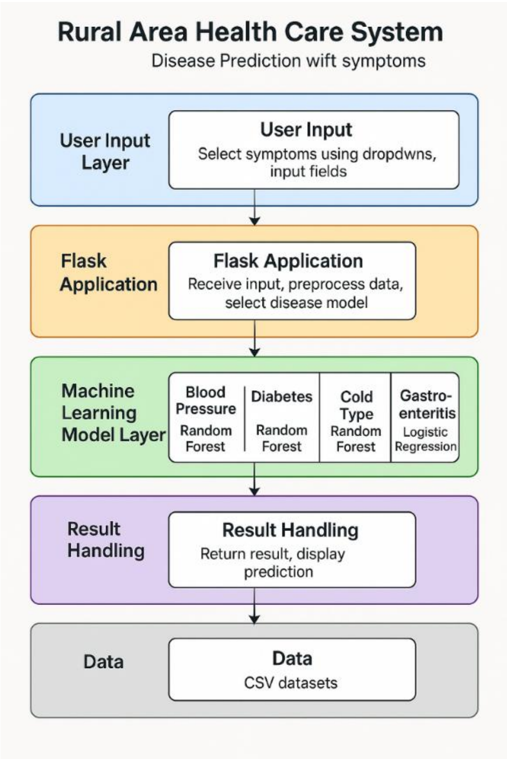

Fig. 1. System Architecture of Rural Health Disease Prediction Model.

The Rural Area Health Care System: Disease Prediction with Symptoms to provide a Web-based, lightweight system to detect disease at early stage in remote areas. It is a modular design with five stages. The front end of the application is the User Interface Layer, written in HTML and styled with CSS, where entered by user symptoms appear as drop downs and forms. These inputs are subsequently sent to the Flask Application Layer which is responsible for routing, input validation and determining the disease model to be invoked. The validated inputs are passed on to the Machine Learning Layer, which contains independent models for all diseases, for example, Random Forest for Asthma, Diabetes, and Blood Pressure, and Logistic Regression for Gastroenteritis. Each model is trained on domain-specific symptom data in the Data Layer from offline CSV files so that the system can operate without internet connectivity. Once a prediction has been generated, the outcome is sent back to the Result Display Layer so that the user is given a definitive diagnosis or health notice. This organized flow provides efficient, offline-capable diagnosis assistance specifically designed for rural healthcare purposes. Fig. 1 shows the system architecture of rural health disease prediction model.

3.2 Data Collection and Preprocessing

For each of the targeted diseases, relevant datasets were collected from public sources and medical references. The diseases and their input features are:

a. Blood Pressure: systolic, diastolic
b. Diabetes: glucose, skin_thickness, insulin, age
c. Cold Type: fever, cough, sore_throat, runny_nose, headache
d. Asthma: age, shortness_of_breath, chest_tightness, coughing, wheezing, family_history, dust_allergy, smoking, cold_air_trigger, exercise_trigger
e. Gastroenteritis: fever, vomiting, diarrhea, abdominal_pain, bloody_diarrhea.

3.3 Disease Prediction Model vs Normal Medical Checkup

Table 1. Differences between Machine Learning Prediction Vs Normal Checkup.

Factor	Disease Prediction Model	Normal Medical Checkup
Speed	Provides instant results after symptom input	Time-consuming: involves multiple tests, waiting for reports
Cost	One-time setup, no per-patient cost once deployed	Repeated expenses for lab tests, doctor fees, and follow-up visits
Accessibility	Can be used offline in remote areas with a web interface	Requires presence of a doctor, clinic, or hospital infrastructure
Reach	Ideal for rural/remote areas	Limited to available healthcare infrastructure
Technology	Uses ML models trained on symptom data	Relies on traditional clinical expertise

Differences between Machine Learning Prediction Vs Normal Checkup has been tabulated in table 1. The Disease Prediction Model provides a huge advantage over conventional medical checkups, particularly in remote and resource-constrained areas. It gives real-time results upon input of symptoms, without requiring time-consuming lab tests and extended waiting times. After deployment, the system becomes cost-effective as it does not involve repeated expenditures per patient, unlike standard checkups which incur consultation fees, diagnostic charges, and follow-up costs. Its availability offline guarantees even remote locations that have poor or no internet connection or infrastructure receive basic healthcare insights. Additionally, the model immensely increases the presence of medical facilities where physicians are scarce. Whilst conventional check-ups rely on expertise in clinics as well as infrastructures, prediction relies on machine learning to deliver uniform, replicable, data-driven diagnoses making it an eminent support tool in rural healthcare.

3.4 Algorithms Used

1) In the rural health care system, various machine learning models were selected for every disease, including prediction, characteristics of input features, and general complexity of data.

a) For predicting blood pressure, Random forest classifier has been used. The model uses two numerical input features: systolic and diastolic blood pressure value. Depending on the input, it predicts whether the user is near normal blood pressure or high blood pressure. The random forest algorithm was chosen because of its effectiveness in dealing with numerical data and offering binary classification with accuracy.

b) For predicting diabetes, random forest classifier was used. It takes glucose levels, skin thickness, insulin levels and age in input. It processes these health indicators to classify the user into diabetes or non-diabetes categories. Random forest was chosen based on its strength and high efficiency in processing multi-facilities therapy.

c) For prediction of cold types, Random forest classifier was used. The model accepts five symptoms-based input: headache, runny nose, sore throat, cough and fever. It predicts one of the different types of cold types, including normal cold, flu or other viral infections. This model was suitable here as it could effectively classify multi-class and take into account the combination of symptoms.

d) For predicting asthma, random forest classifier was chosen as the disease depends on many binary (yes/no) indicators. The model uses the following input features: age, shortness of breath, chest tightness, cough, wheezing, family history, dust allergy, smoking, cold air trigger and exercise trigger. Based on these inputs, the model predicts whether the user is likely to have asthma. Random forest was ideal due to many categories of characteristics and their ability to handle the interaction.

e) For predicting gastroenteritis, logistic regression was used. These five symptoms take input: fever, vomiting, diarrhea, abdominal pain and bloody diarrhea. Depending on these symptoms, it predicts whether the user is infected with gastroenteritis. Logistic regression was chosen for effectiveness with simplicity, rapid performance and direct symptomatic patterns in binary classification functions. Each model was individually trained using the specific dataset and was later integrated into the web application using the flask, which allows for real-time predictions based on the user input.

2) The corresponding models were chosen for handling the data types and prediction needs of each disease. We selected Random Forest Classifier as it has high accuracy and is immune to overfitting, and it works well with both categorical and numerical data. One natural fit is for health, where symptoms might interact in complex manners; it specializes for the prediction of ailments like asthma, diabetes, and other forms of the cold. The ensemble aspect enables to catch these relations and return consistent results even for small datasets. On the contrary, for prediction of gastroenteritis we chose Logistic Regression because of its simplicity and capacity to quickly predict in binary classification. *It works well when the relationship between input symptoms and the output is linear, and it requires fewer computational resources, making it suitable for a lightweight, offline-capable system intended for rural healthcare settings. [15]*

a) Logistic Regression (LR)

- A probabilistic model that predicts the likelihood of a disease based on the symptoms:

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn)}} \tag{1}$$

- **Objective:** Minimize the logistic loss function to determine the optimal weights β. [16]

b) Random Forest (RF)

An ensemble of decision trees to improve robustness:

- Combines predictions from multiple trees using majority voting.
- Reduces overfitting compared to single decision tree.

Table 2. Disease Prediction Accuracy.

Disease Name	Accuracy
Blood Pressure	1.00
Asthma	0.88
Gastroenteritis	0.72
Cold type	1.00
Diabetes	0.71

Table 2 shows the disease prediction accuracy.

3) Use Case :

The primary use case of the Rural Health Care System is to provide preliminary disease diagnosis in remote and underserved areas where access to doctors and diagnostic facilities is limited. In such regions, healthcare centers often lack expert medical staff but may still have access to basic devices like computers or mobile tablets.

This system enables health workers or patients themselves to input symptoms and receive immediate predictions for common health conditions such as blood pressure anomalies, diabetes, asthma, cold types, and gastroenteritis. For example, a rural clinic worker can input a patient's glucose level and age to quickly assess diabetes risk or use symptom checklists to detect asthma or cold-related issues in children. Since the application runs on a lightweight web interface and requires no internet once deployed, it is highly suitable for offline, low-resource environments. This makes it especially beneficial for health camps, schools, mobile clinics, and primary health centers in villages, where real-time expert advice is often unavailable.

Overall, the system acts as a first line screening tool, helping prioritize patients who need immediate attention and reducing the burden on rural healthcare infrastructure.

4) Advantages :

The application has several advantages particularly for rural and deprived populations. It enables quick and accurate prediction of diseases with input symptoms using machine learning without the need of expensive diagnostic tests, and saving time. It can be used offline so you can read even when you have no or limited internet access (such as in remote areas where internet access is not always available). The procedure is cost-effective as it prevents regular repeated clinical visits and test charges and thus is affordable for low-income people. The system is very easy to use and can be operated by healthcare professionals who may not necessarily have a high degree of technical knowledge due to the intuitive web interface. Data storage for patient history, multilingual support for wider penetration, medicine recommendation with precautions and an in-app chatbot for solving health related queries is also available. These features combined mean it offers an affordable, scalable and successful means of improving healthcare in rural regions.

5) Feature Engineering:

The feature extraction method applied in the Rural Health Care System. When the user uploads symptom-testing data on the web interface, a program execution is initiated. These inputs are then classified into numerical or categorical. Numerical variables such as age, systolic blood pressure, and glucose levels do not undergo categorization. Symptoms such as cough, fever, and environmental factors are relabeled as binary variables using "1" for present and "0" for absent. Once all inputs are processed, they are combined into a unified feature vector. This vector is passed to the disease-specific machine learning model either Random Forest or Logistic Regression depending on the selected disease. Finally, the model generates a prediction based on the processed inputs, which is then displayed to the user. Fig. 2 depicts the flow diagram of health care disease prediction system.

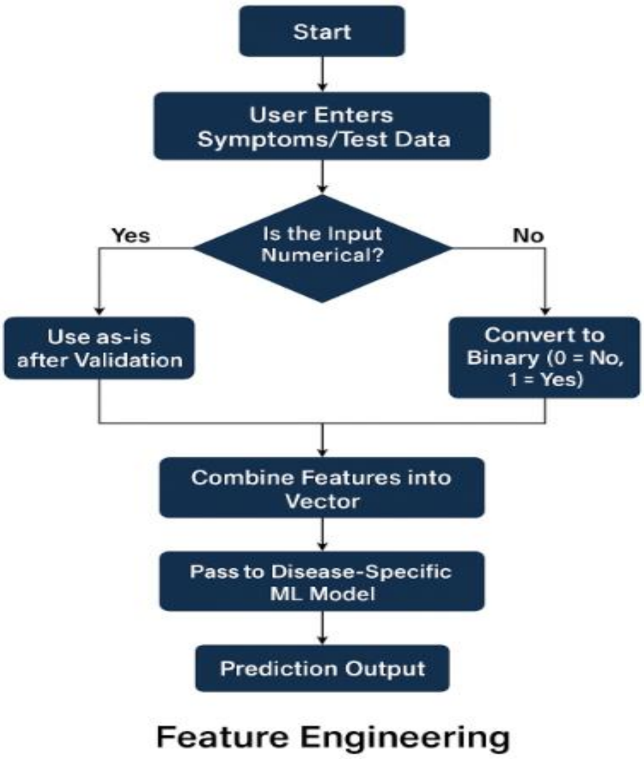

Fig. 2. Flow Diagram of Health Care Disease Prediction System.

4 Future Scope & Enhancements

To make the system more robust, scalable, and community-friendly, the following enhancements are proposed:

- *Multi-Disease Expansion :* The system can be extended to include predictions for other popular rural diseases such as malaria, typhoid, tuberculosis, or anemia using relevant symptoms and updated datasets. This modularity will allow continuous scaling as health data increases.
- *Login & Record Management :* Adding login-based access to the workers of the Rural Health Center will allow them to store, recover and update the patient records, including symptoms, diagnosis history and treatment details. This data will be necessary for tracking health trends, repeating symptoms and follow -up care.
- *Medicine Suggestion & Precaution Module:* A new section named "Medicines" can be introduced to provide AI recommended, over-the-counter medicine names (for initial guidance) and general precautionary steps for disease. This will help patients take early actions even before consulting a doctor.
- *Multilingual Support :* Language barriers are a major issue in rural communities. A language translation feature can enable users to switch the application's interface and outputs into regional languages, ensuring that the platform is accessible and comprehensible to all users, regardless of literacy level in English.
- *AI-Powered Chatbot Assistant :* A chatbot facility can be incorporated as an in-built feature that enables users to communicate with the system in a natural way by asking questions or clearing doubts regarding symptoms, diseases, drugs, and preventive measures. The virtual assistant would be online 24/7, enhancing user interaction and trust, particularly where real-time medical assistance is not always readily available.

5 Results

The proposed Rural Health Care System was evaluated for its prediction accuracy across five major diseases using symptom-based and test-based datasets. The system demonstrated promising results, highlighting its effectiveness in early diagnosis and support for rural healthcare needs.

The Random Forest Classifier, applied to Blood Pressure, Asthma, Cold Type, and Diabetes, showed high robustness in handling both binary and numerical features. It achieved 100% accuracy for Blood Pressure and Cold Type, indicating excellent model performance when symptoms are distinct and well-defined. Asthma prediction reached an accuracy of 88%, effectively capturing the impact of both environmental triggers and patient history. Diabetes prediction yielded an accuracy of 71%, showing decent performance with basic health metrics like glucose, insulin, and age.

The Logistic Regression model, used for Gastroenteritis, achieved 72% accuracy, making it a lightweight yet effective classifier for binary symptom patterns. While slightly lower than other models, it still provides practical diagnostic support in low-resource settings.

These results confirm that the selected models are suitable for their respective diseases, balancing accuracy, speed, and usability. The system delivers fast predictions and is well-suited for offline use in rural clinics, with potential for future model optimization and disease expansion. Fig. 3 shows the Home Interface of Rural Health Care Prediction System.

Fig. 3. Home Interface of Rural Health Care Prediction System.

6 Conclusion

The rural healthcare system uses machine intelligence to enable quick and accurate prediction of the diseases profiled with symptoms entered in real time, serving up vital support in low-resource areas. With support for offline usage, and services such as patient's history data tracking and AI-powered consultation, the tool increases efficiency in diagnostics and empowers access to healthcare services. This could be a game-changer for health in the country - helping rural health workers to diagnose and treat diseases at the community level.

References

[1] X. Chen, Y. Liu, and W. Zhao, "A Machine Learning-Based Framework for Accurate Disease Diagnosis in Rural Areas," Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 6657319, 2021. DOI: 10.1155/2021/6657319.

[2] S. Kumar and R. Rajasekaran, "A Lightweight Decision Support System for Rural Diagnostics Using Logistic Regression," Procedia Computer Science, vol. 171, pp. 1914–1921, 2020. DOI: 10.1016/j.procs.2020.04.205.

[3] A. Reddy, M. S. Reddy, and T. R. Rao, "Symptom-based Disease Prediction using Machine Learning," Journal of Medical Systems, vol. 45, no. 7, p. 62, 2021. DOI: 10.1007/s10916-021-01747-9.

[4] A. Prakash and K. Rani, "A Symptom-based ML System for Primary Healthcare," Health Informatics Journal, vol. 28, no. 1, pp. 146–156, 2022. DOI: 10.1177/14604582211066398.

[5] N. Sharma, P. Verma, and S. Saxena, "A Decision Tree and Ensemble-Based Framework for Symptom-Oriented Healthcare Diagnosis," IEEE Access, vol. 8, pp. 112675–112685, 2020. DOI: 10.1109/ACCESS.2020.3002034.

[6] P. Singh and J. Kaur, "Early-Stage Diabetes Detection Using Random Forest in Low-Resource Environments," International Journal of Health Sciences and Research, vol. 9, no. 12, pp. 120–125, 2019.

[7] R. Gupta, M. Patel, and N. Sharma, "Cold Type Classification Using Random Forest and Gradient Boosting," Biomedical Signal Processing and Control, vol. 62, 2020, Article ID 102108. DOI: 10.1016/j.bspc.2020.102108.

[8] S. Patil and S. Kulkarni, “An Offline Rural Diagnostic Assistant Using Machine Learning and CSV Data,” Journal of Information Technology and Software Engineering, vol. 11, no. 3, 2021. DOI: 10.35248/2165-7866.21.11.284.
[9] Y. Zhang, L. Wu, and T. Lin, “Lightweight and Offline-Capable ML for Rural Health,” in Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC), 2020, pp. 1–6. DOI: 10.1109/GHTC46280.2020.9342899.
[10] S. Rao and D. Kumar, “Integration of Random Forest and Logistic Regression in Scalable Disease Diagnosis Systems,” International Journal of Medical Informatics, vol. 150, 2021, Article ID 104429. DOI: 10.1016/j.ijmedinf.2021.104429.
[11] A. Ahmed, H. N. Khan, and S. Farooq, “Disease Prediction from Symptoms Using Ensemble Learning,” Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 5, pp. 1578–1586, 2021. DOI: 10.1016/j.jksuci.2021.03.006.
[12] V. Mishra and A. Das, “Scalable Machine Learning Model for Rural Disease Prediction Using Open Datasets,” Procedia Computer Science, vol. 167, pp. 2234–2241, 2019. DOI: 10.1016/j.procs.2020.03.273.
[13] A. Nair, P. Menon, and R. Thomas, “Fever Diagnostic Assistant: A Non-Predictive Informative Module for Rural Healthcare,” IEEE Access, vol. 9, pp. 110024–110031, 2021. DOI: 10.1109/ACCESS.2021.3102579.
[14] D. Basak and A. Ghosh, “Multilingual Support in Rural Healthcare Applications: A Usability Study,” Journal of Biomedical Informatics, vol. 110, 2020, Article ID 103573. DOI: 10.1016/j.jbi.2020.103573.
[15] F. Costa, M. Amorim, and C. de Souza, "Machine Learning for Predicting Chagas Disease Infection in Rural Areas of Brazil," IEEE Latin America Transactions, vol. 20, no. 6, pp. 1030–1036, June 2022. DOI: 10.1109/TLA.2022.9865421.
[16] T. S. Rao and M. Narayan, "Utilizing Logistic Regression for Disease Prediction in Primary Healthcare Centers," Indian Journal of Public Health Research & Development, vol. 12, no. 1, pp. 299–304, 2021. DOI: 10.37506/ijphrd.v12i1.13791.