Integrated Machine Learning Framework for Multi Disease Prediction and Geolocation-Based Healthcare Recommendations

ARSALAN[1], SHURAIM SHAKEEL BHAT[2], B P CHANDANA[3], REYA JAVAID[4]
[1, 2, 3, 4]Department of CSE, Ghousia College of Engineering, Ramanagaram, Karnataka, India

Abstract- Chronic diseases such as diabetes, heart disease, breast cancer, and diabetic retinopathy continue to impose a substantial burden on healthcare systems due to delayed diagnosis and limited post-diagnostic guidance. Many existing digital health platforms focus solely on prediction while neglecting actionable follow-up, such as identifying suitable healthcare providers. This paper presents an integrated web-based machine learning framework that performs multi-disease risk prediction and provides location-based healthcare recommendations. The system supports four disease models: Logistic Regression for diabetes and diabetic retinopathy, and Random Forest classifiers for heart disease and breast cancer. Users manually input clinical parameters, which are preprocessed and evaluated using pre-trained models deployed on a centralized server. Experimental evaluation on publicly available datasets demonstrates classification accuracies of 75.32% for diabetes, 99.50% for diabetic retinopathy, 90.16% for heart disease, and 83.23% for breast cancer. Beyond prediction, the framework incorporates a geolocation module that recommends nearby hospitals and specialists based on the predicted outcome. The results indicate that combining disease prediction with post-prediction guidance improves practical usability, although clinical deployment would require validation on real-world patient data.

Keywords—Disease Prediction, Machine Learning, Healthcare Recommendation, Logistic Regression, Random Forest, Web-Based Healthcare System

I. INTRODUCTION

Non-communicable diseases account for a growing proportion of global morbidity and mortality. Diabetes and cardiovascular diseases are among the leading causes of long-term complications, while breast cancer remains a primary contributor to cancer-related deaths in women. Diabetic retinopathy, although preventable, is a major cause of avoidable blindness when early detection is missed.

Machine learning techniques have been widely applied to individual disease prediction tasks. However, most deployed systems are disease-specific and operate in isolation. Moreover, prediction outputs are often binary labels without contextual guidance, leaving patients uncertain about subsequent actions.

This work addresses two gaps. First, it integrates multiple disease prediction models into a single web application. Second, it extends prediction results with geolocation-based hospital and doctor recommendations. The primary contributions of this paper are:

- A unified architecture supporting four disease prediction models.
- Algorithm selection tailored to dataset characteristics.
- Integration of prediction outputs with healthcare recommendation logic.
- Experimental evaluation using standard accuracy metrics.

II. PROPOSED METHODOLOGY

2.1. System Architecture

The system follows a client–server architecture consisting of frontend, backend, and data/model layers. User requests are processed synchronously, and prediction results along with recommendations are returned for visualization.

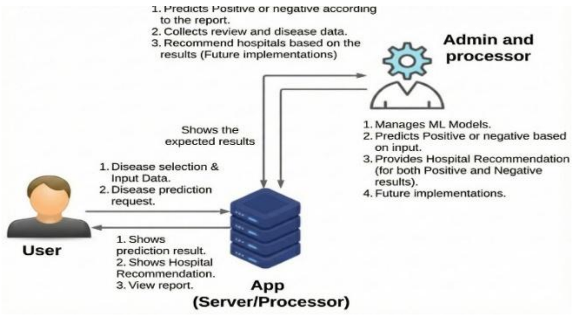
1. Predicts Positive or negative according to the report.
2. Collects review and disease data.
3. Recommend hospitals based on the results (Future implementations)
Admin and processor
1. Manages ML Models.
2. Predicts Positive or negative based on input.
3. Provides Hospital Recommendation (for both Positive and Negative results).
4. Future implementations.
Shows the expected results
1. Disease selection & Input Data.
2. Disease prediction request.
User
1. Shows prediction result.
2. Shows Hospital Recommendation.
3. View report.
App (Server/Processor)

Fig.1.System architecture

2.2. Data Preprocessing

Each dataset undergoes preprocessing prior to training and inference. Numeric attributes are converted to appropriate data types, missing values are handled using mean imputation, and feature scaling is applied using standard normalization. For diabetes prediction, Pearson correlation analysis was used to select glucose and body mass index as dominant predictors.

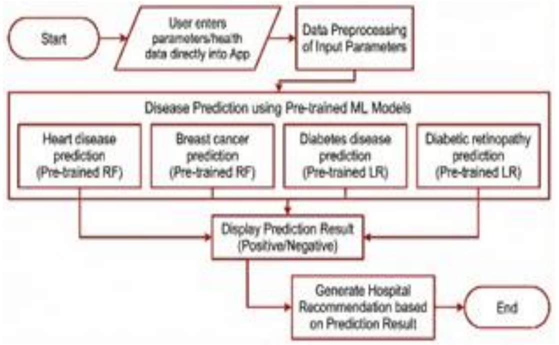
Start
User enters parameters/health data directly into App
Data Preprocessing of Input Parameters
Disease Prediction using Pre-trained ML Models
Heart disease prediction (Pre-trained RF)
Breast cancer prediction (Pre-trained RF)
Diabetes disease prediction (Pre-trained LR)
Diabetic retinopathy prediction (Pre-trained LR)
Display Prediction Result (Positive/Negative)
Generate Hospital Recommendation based on Prediction Result
End

Fig .2. Flow chart of Data Preprocessing

2.3 Algorithm Selection Rationale

Algorithm choice was driven by dataset size, feature characteristics, and interpretability requirements:

- Logistic Regression (Diabetes, Diabetic Retinopathy):
  Logistic Regression was selected due to its suitability for binary classification and its ability to provide probabilistic outputs. In diabetes prediction, interpretability of coefficients is important for understanding the contribution of clinical features. For retinopathy, the dataset exhibited near-linear separability after preprocessing, making Logistic Regression sufficient without introducing unnecessary model complexity.
- Random Forest (Heart Disease, Breast Cancer): Random Forest classifiers were employed to handle nonlinear feature interactions and reduce overfitting. These datasets contain multiple correlated attributes, and ensemble learning improves robustness compared to single decision trees.

This selection prioritizes stability over novelty. No deep learning models were used, as the dataset sizes and feature structures did not justify higher model complexity.

2.4 Geolocation and Recommendation Logic

Following a positive prediction, the system queries a hospital database containing geographic coordinates and specialization metadata. Nearby hospitals are identified based on distance thresholds and filtered by disease specialization.

III. EXPERIMENTAL RESULTS

Model performance was evaluated using classification accuracy on held-out test sets. The results are summarized in Table I.

Disease	Algorithm	Accuracy (%)
Diabetes	Logistic Regression	75.32
Diabetic Retinopathy	Logistic Regression	99.50
Heart Disease	Random Forest	90.16
Breast Cancer	Random Forest	83.23

Fig.3. Accuracy Table

The retinopathy model exhibits unusually high accuracy, which likely reflects dataset bias or class imbalance rather than true clinical robustness.

IV. DISCUSSION

The results indicate that classical machine learning models remain competitive for structured clinical datasets. Random Forest classifiers consistently outperformed Logistic Regression on higher-dimensional datasets, while Logistic Regression remained adequate for simpler binary classification tasks.

However, the system relies on manually entered data, which introduces user-dependent noise. Additionally, the recommendation module is advisory in nature and does not account for hospital capacity, physician availability, or patient history. These limitations restrict direct clinical deployment

V. CONCLUSION AND FUTURE WORK

This paper presented an integrated machine learning framework for multi-disease prediction combined with geolocation-based healthcare recommendations. The system demonstrates that coupling prediction with post-diagnostic guidance improves practical relevance compared to standalone classifiers.

Future work will focus on:

- Expanding disease coverage.
- Incorporating real-time sensor and electronic health record data.
- Applying cost-sensitive and recall-oriented evaluation metrics.
- Validating the system using clinically curated datasets.
- Extending the platform to mobile environments.

Without clinical validation, the system should be viewed strictly as a decision-support prototype rather than a diagnostic tool.

ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance and technical support provided by the faculty of the Department of Computer Science and Engineering, Ghousia College of Engineering, throughout the development of this work. Their constructive feedback and academic supervision were instrumental in ensuring the completeness and rigor of this research

REFERANCES

[1] Krishnamoorthi, R., Joshi, S., AL Marzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, *2022*.

[2] Rayapu, L., Chakraborty, K., & Valluru, L. (2021). Marine algae as a potential source for anti-diabetic compounds-A brief review. *Current Pharmaceutical Design*, *27*(6), 789-801.

[3]Bhat, S. S., Selvam, V., Ansari, G. A., Ansari, M. D., & Rahman, M. H. (2022). Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district Bandipora. *Computational Intelligence and Neuroscience*, *2022*.

[4] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, *10*(1), 11981.

[5] Greenwald, H. D., Kennedy, L. C., Hinkle, A., Whitney, O. N., Fan, V. B., Crits-Christoph, A., ... & Nelson, K. L. (2021). Tools for interpretation of wastewater SARS-CoV-2 temporal and spatial trends demonstrated with data collected in the San Francisco Bay Area. *Water research X*, *12*, 100111.

[6] Shafi, S., & Ansari, G. A. (2021, May). Early prediction of diabetes disease & classification of algorithms using machine learning approach. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

[7] Rout, M., & Kaur, A. (2020, June). Prediction of diabetes risk based on machine learning techniques. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 246-251). IEEE.

[8] Bell, J. (2022). What is machine learning? *Machine Learning and the City: Applications in Architecture and Urban Design*, 207-216.

[9] Bhat, S. S., Banu, M., Ansari, G. A., & Selvam, V. (2023). Diabetes detection system using machine learning algorithms. International Journal of Electronic Healthcare, 13(3), 231-246.

[10] Qawqzeh, Y. K., Bajahzar, A. S., Jemmali, M., Otoom, M. M., & Thaljaoui, A. (2020). Classification of diabetes using photoplethysmogram (PPG) waveform analysis:

Logistic regression modeling. *BioMed Research International*, *2020*.

[11] Annamalai, R., & Nedunchelian, R. (2021). Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. *Computational Intelligence and Neuroscience*, *2021*.

[12] Ihnaini, B., Khan, M. A., Khan, T. A., Abbas, S., Daoud, M. S., Ahmad, M., & Khan, M. A. (2021). A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning. *Computational Intelligence and Neuroscience*, *2021*.

[13] Vives-Boix, V., & Ruiz-Fernandez, D. (2021). Diabetic retinopathy detection through convolutional neural networks with synaptic metaplasticity. *Computer Methods and Programs in Biomedicine*, *206*, 106094.

[14] Tsarapatsani, K., Sakellarios, A. I., Pezoulas, V. C., Tsakanikas, V. D., Kleber, M. E., März, W., ... & Fotiadis, D. I. (2022, July). Machine Learning Models for Cardiovascular Disease Events Prediction. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1066-1069). IEEE.

[15] Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3204-3225.

[16] Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *Journal of Infection and Public Health*, *13*(9), 1274-1289.

[17] Malik, S., Harous, S., & El-Sayed, H. (2021). Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. In *Modelling and Implementation of Complex Systems: Proceedings of the 6th International Symposium, MISC 2020, Batna, Algeria, October 24-26, 2020 6* (pp. 95-106). Springer International Publishing.

[18] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179.

[19] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

[20] Zhang, G. (2016). Making Sense of Conflict in Distributed Teams: A Design Science Approach.