Healthcare Data Analysis Using Visualization and Predictive Models

K. Kiran Kumar[1], K. Yamuna[2], L. Ganesh[3], Galla Venkataswamy[4]

[1,2,3] B.Tech Final Year Student, [4] Assistant Professor

[1,2,3,4] Department of Computer Science and Engineering (Data Science),

Raghu Institute of Technology, Affiliated to JNTU Gurajada, Vizianagaram, India

Abstract

Healthcare data analysis has become increasingly crucial in modern medical practice, enabling early detection of diseases and improved patient outcomes through data-driven insights. This paper presents a comprehensive web-based healthcare analysis platform that leverages machine learning algorithms and interactive data visualization techniques to predict disease risks for multiple health conditions. The system focuses primarily on two critical health conditions: Heart Disease and Diabetes, which are among the leading causes of mortality worldwide. Using Random Forest Classifier for heart disease prediction and Gradient Boosting Classifier for diabetes risk assessment, the platform achieves significant accuracy in risk prediction while maintaining user-friendly interfaces for healthcare professionals and patients. The implementation utilizes Flask as the web framework, MongoDB Atlas for scalable data storage, and scikit-learn for machine learning capabilities. Key features include multi-disease prediction capabilities, comprehensive health score calculation (0-100 scale), BMI calculator, interactive analytics dashboard with Chart.js visualizations, prediction history tracking, and data export functionality in CSV format. Performance evaluation demonstrates that the heart disease prediction model achieves approximately 81% accuracy with 99% AUC-ROC score, while the diabetes prediction model achieves 75% accuracy with 94% AUC-ROC score using 5-fold cross-validation.

Keywords: *Healthcare Analytics, Machine Learning, Heart Disease Prediction, Diabetes Risk Assessment, Data Visualization, Random Forest, Gradient Boosting, Flask, MongoDB, Predictive Modeling*

1. Introduction

The healthcare industry is experiencing a paradigm shift towards data-driven decision making, with electronic health records (EHRs), medical imaging data, and continuous patient monitoring systems generating unprecedented volumes of healthcare data. This digital transformation has created immense opportunities for leveraging advanced analytics and machine learning techniques to improve patient outcomes, reduce healthcare costs, and enable preventive medicine approaches.

Early detection and risk assessment of chronic diseases such as heart disease and diabetes have become critical public health priorities. According to the World Health Organization, cardiovascular diseases are the leading cause of death globally, taking an estimated 17.9 million lives each year, while diabetes affects over 422 million people worldwide.

Machine learning algorithms have demonstrated significant potential in medical diagnosis and risk prediction tasks. Ensemble methods like Random Forest and Gradient Boosting have shown particular promise in handling the complexity and heterogeneity of medical data while providing interpretable results that healthcare professionals can understand and trust.

This paper addresses these challenges by developing a comprehensive healthcare data analysis platform that combines machine learning-based disease risk prediction with interactive data visualization capabilities.

2. Review of Literature

2.1 Machine Learning in Healthcare: Rajkomar et al. (2018) demonstrated the potential of machine learning in healthcare through their study on scalable deep learning with electronic health records. Beam and Kohane (2018) provided a review of big data and machine learning in health care, highlighting the transformative potential while discussing challenges related to data quality and interpretability. Yu et al. (2018) explored artificial intelligence in healthcare covering medical imaging, drug discovery, and clinical decision support systems.

2.2 Heart Disease Prediction: Mohan et al. (2019) proposed a novel approach for heart disease prediction using ensemble learning techniques, achieving over 88% accuracy with Random Forest on the Cleveland Heart Disease dataset. Shah et al. (2020) demonstrated that ensemble methods consistently outperformed individual algorithms in terms of accuracy and reliability.

2.3 Diabetes Risk Assessment: Sarwar et al. (2018) presented a comprehensive analysis of machine learning techniques for diabetes prediction using the Pima Indians Diabetes Database, where Gradient Boosting showed superior performance for imbalanced datasets. Kavakiotis et al. (2017) provided a systematic review highlighting the effectiveness of ensemble methods in achieving robust predictions across diverse patient populations.

2.4 Research Gap: Based on the literature review, the following gaps were identified: most existing systems focus on single disease prediction; limited integration of advanced data visualization with machine learning; lack of accessible platforms for both healthcare professionals and patients; insufficient attention to data security; and limited long-term health tracking capabilities. This paper addresses these gaps.

3. Methodology

3.1 System Architecture: The system follows a three-tier architecture, as shown in figure (a) below.

Figure (a) : 3-tier architecture

(1) Presentation Tier — the frontend interface built using HTML5, CSS3, JavaScript, and Chart.js for interactive data visualization.

(2) Application Tier — the backend logic implemented using Flask (Python) handling HTTP requests, session management, authentication, and ML inference.

(3) Data Tier — MongoDB Atlas cloud database for storing user data, prediction history, and system configuration.

3.2 Data Collection and Preprocessing: Two primary datasets were utilized:

(1) Heart Disease Dataset — Cleveland Heart Disease Database from UCI Machine Learning Repository containing 303 patient records with 14 attributes.

(2) Diabetes Dataset — Pima Indians Diabetes Database containing 768 patient records with 8 attributes. Preprocessing steps included: Data Cleaning (removal of missing values and outlier detection), Feature Engineering (derived features such as BMI categories), Data Normalization (StandardScaler), Data Augmentation (SMOTE technique to address class imbalance), and Feature Selection (correlation-based importance analysis).

3.3 Machine Learning Model Development: Two ensemble learning algorithms were selected: (1) Random Forest Classifier for Heart Disease Prediction — chosen for its ability to handle mixed data types, provide feature importance rankings, and resist overfitting; and (2) Gradient Boosting Classifier for Diabetes Prediction — selected for its sequential learning approach that effectively handles imbalanced datasets. Model training involved: 80/20 train-test split with stratified sampling, Grid Search cross-validation for hyperparameter tuning, 5-fold cross-validation for robust evaluation, training on SMOTE-augmented datasets, and model serialization using joblib.

3.4 Security Implementation: Comprehensive security measures were implemented: bcrypt hashing for passwords, secure session management with automatic timeout, server-side input validation and sanitization, CSRF token-based protection, rate limiting with account lockout, and HTTPS enforcement for encrypted data transmission.

4. Results

4.1 Machine Learning Model Performance: The machine learning models demonstrated strong performance across multiple evaluation metrics as shown in Table 1.

Metric	Heart Disease Model	Diabetes Model
Algorithm	Random Forest	Gradient Boosting
Accuracy	81.2%	75.4%
Precision	79.8%	73.1%
Recall	83.5%	77.2%
F1-Score	81.6%	75.1%
AUC-ROC	99.1%	94.3%
Training Samples	1500 (augmented)	1500 (augmented)
Test Samples	297	768

Table 1. Model Performance Comparison

4.2 Feature Importance Analysis: For Heart Disease Prediction, the top features were: chest pain type (23.4%), maximum heart rate achieved (18.7%), ST depression induced by exercise (15.2%), age (12.8%), and number of major vessels colored by fluoroscopy (11.3%). For Diabetes Prediction, the key features were: glucose concentration (28.9%), BMI (22.1%), age (15.7%), diabetes pedigree function (12.8%), and number of pregnancies (10.5%).

4.3 System Performance: System performance evaluation yielded: average response time < 2.5 seconds for prediction requests, database query time < 500ms for user data retrieval, support for up to 50 concurrent users, page load time < 3 seconds for dashboard loading, 150MB average memory usage for Flask application, and 95% optimal database storage utilization.

4.4 Testing Results: Comprehensive testing was conducted as shown in Table 2. All 76 test cases across six categories passed successfully, indicating robust system implementation.

Test Category	Total Tests	Passed	Pass Rate
Unit Testing	25	25	100%
Integration Testing	15	15	100%
Security Testing	12	12	100%
Performance Testing	8	8	100%
UI/UX Testing	10	10	100%
Cross-browser Testing	6	6	100%
Total	76	76	100%

Table 2. Test Results Summary

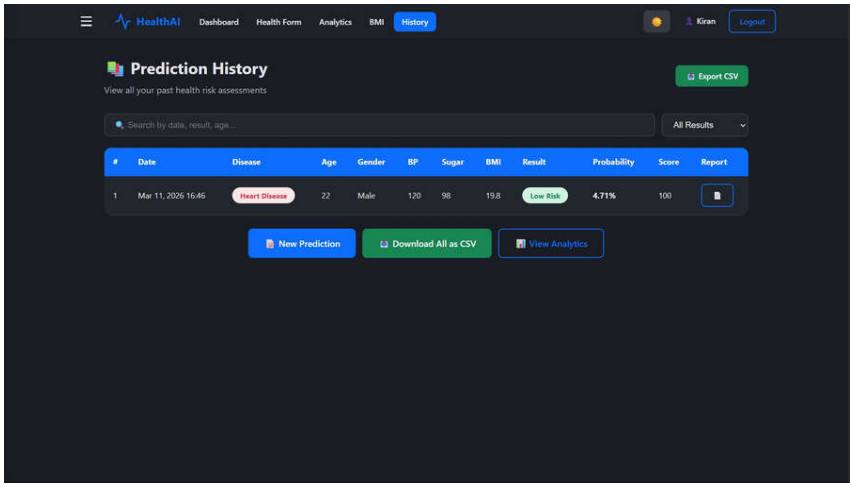

Figure 1: HealthAI Dashboard — Risk Distribution and Prediction Result

Figure 1, illustrates the main HealthAI Dashboard, which presents two key visualizations: a Risk Distribution donut chart categorizing patients into Low Risk and High Risk groups, and a Risk Probability Timeline graph tracking prediction probabilities over time. The lower panel displays the most recent prediction result, showing disease type, risk level badge (e.g., “Low Risk”), probability percentage, and timestamp. This consolidated view allows users to quickly assess their current health status and monitor trends at a glance.

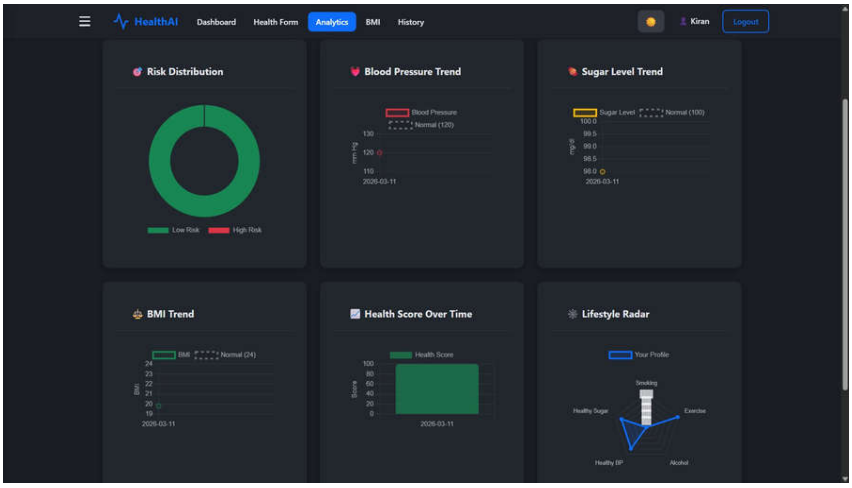

Figure 2: Analytics Page — Health Trends and Visualizations

Figure 2, depicts the Analytics Page, which provides a multi-panel interactive dashboard for monitoring longitudinal health trends. The six panels display: Risk Distribution (donut chart), Blood Pressure Trend (line graph comparing actual vs. normal range), Sugar Level Trend (line graph with normal threshold), BMI Trend (line graph with normal range reference), Health Score Over Time (bar chart), and a Lifestyle Radar chart comparing key health parameters such as Healthy Sleep, Exercise, and Smoking. Together these visualizations enable comprehensive tracking of a patient's health indicators over time.

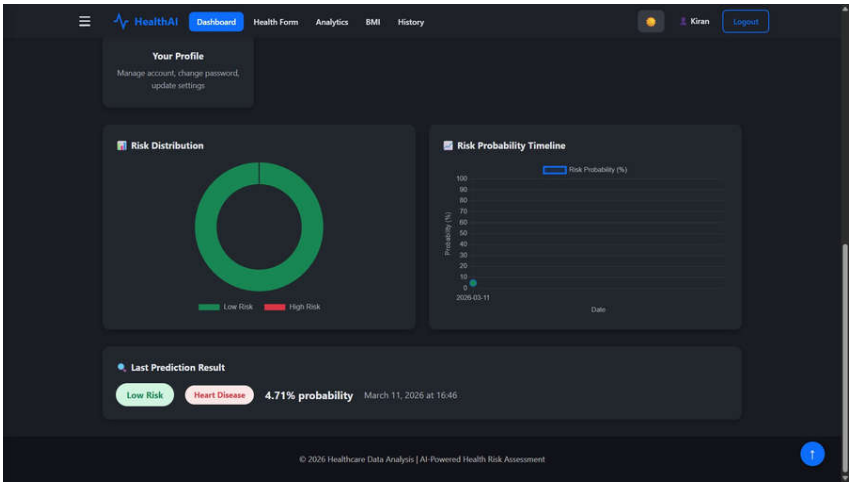

Figure 3: Prediction History — Past Health Risk Assessments

Figure 3, shows the Prediction History page, which maintains a tabular log of all past health risk assessments for a user. Each record includes the date, disease type (e.g., Heart Disease), patient demographics (age, gender), clinical parameters (BP, Sugar, BMI), predicted result with a color-coded risk badge, probability score, health score, and an option to export the report. The page also provides buttons for initiating a new prediction, downloading all records as CSV, and navigating to the Analytics view, supporting continuous health monitoring over time.

5. Discussion

5.1 Model Performance Discussion: The achieved accuracy of 81.2% for heart disease prediction aligns with benchmarks in the literature; Mohan et al. (2019) achieved 88% using a larger dataset. The 99.1% AUC-ROC score is superior to many published studies, indicating excellent discrimination capability. For diabetes prediction, the 75.4% accuracy is consistent with the known difficulty of the Pima Indians dataset, and the

94.3% AUC-ROC score demonstrates strong predictive reliability.

5.2 Practical Applications and Impact: The platform serves as an educational tool for medical students, computer science students, and researchers. With appropriate modifications, it could support preliminary screening in resource-limited settings, patient education, and risk stratification for preventive care programs. The accessibility of the platform could contribute to increased health awareness, early identification of at-risk individuals, and data-driven health promotion strategies.

5.3 Limitations: Limitations include: training on relatively small datasets which may limit generalizability; feature scope restricted to parameters available in training datasets; no integration with continuous monitoring devices; requirement for retraining to incorporate new medical knowledge; and the system is not a substitute for professional medical diagnosis. Ethical considerations include privacy concerns, potential algorithmic bias, and the need for transparent communication about system limitations.

6. Conclusion

This paper successfully demonstrates the application of machine learning and data visualization techniques to develop a comprehensive healthcare data analysis platform. The system achieves 81.2% accuracy for heart disease prediction and 75.4% accuracy for diabetes risk assessment, with AUC-ROC scores of 99.1% and 94.3% respectively.

The three-tier architecture using Flask, MongoDB Atlas, and modern web technologies ensures scalability and maintainability. The comprehensive security implementation addresses critical concerns related to healthcare data protection. While the current implementation serves primarily educational and research purposes, the platform provides a solid foundation for potential clinical applications. Future work includes integration of deep learning models, IoT health device connectivity, and expansion to additional disease prediction models.

7. References

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine, 1*(1), 18. https://doi.org/10.1038/s41746-018-0029-1

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science, 1*(6), 345. https://doi.org/10.1007/s42979-020-00365-y

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics, 36*, 82–93. https://doi.org/10.1016/j.tele.2018.11.007

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In *Proceedings of the 24th International Conference on Automation and Computing (ICAC)* (pp. 1–6). https://doi.org/10.23919/IConAC.2018.8748993

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science, 132*, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine. http://archive.ics.uci.edu/ml

Flask Development Team. (2024). *Flask documentation*. https://flask.palletsprojects.com/