Logistic Regression for Early Heart Disease Detection: Improving E-Healthcare through Machine Learning

Omar Farooq, Chiara Bianchi

University of Dhaka, Bangladesh

University of Bologna, Italy

Abstract:

Heart disease remains one of the leading causes of mortality worldwide. Early detection is critical in improving patient outcomes and reducing healthcare costs. In this paper, we explore the application of logistic regression in predicting heart disease risk, leveraging patient data such as blood pressure, cholesterol levels, and lifestyle factors. Logistic regression is a powerful yet interpretable machine learning model well-suited for binary classification tasks, such as identifying whether a patient is at risk of developing heart disease. The aim of this research is to demonstrate how logistic regression can be effectively implemented in e-healthcare systems to enhance early diagnosis, improve personalized treatment plans, and ultimately reduce mortality rates.

Keywords: Logistic Regression, Heart Disease, Early Detection, Machine Learning, E-Healthcare, Predictive Modeling, Healthcare Data

I. Introduction:

Heart disease is responsible for a significant portion of global deaths, and early detection can drastically improve survival rates. As healthcare systems move towards digitization, the integration of machine learning models offers an opportunity to enhance diagnostic accuracy and personalize patient care. Logistic regression, a widely used statistical model, excels in binary classification tasks and can be pivotal in predicting whether patients are at risk of developing heart disease based on a variety of input features, such as age, gender, cholesterol levels, and lifestyle habits[1]. E-healthcare platforms can leverage such machine learning models to provide timely interventions, making medical care more accessible, especially in regions with limited healthcare infrastructure. In this research, we delve into how logistic regression can be effectively used to predict early signs of heart disease and how this integration could transform e-healthcare delivery.

Heart disease, encompassing conditions such as coronary artery disease, arrhythmias, and heart failure, is a leading cause of death globally. The World Health Organization (WHO) reports that cardiovascular diseases (CVDs) claim approximately 17.9 million lives each year, accounting for 31% of all global deaths. Factors like sedentary lifestyles, unhealthy diets, smoking, and genetic predisposition contribute to the rising prevalence of heart disease. Early detection is crucial for effective treatment and prevention, as timely interventions can reduce the risk of severe complications, improve quality of life, and decrease mortality rates[2].

Traditional methods of diagnosing heart disease rely heavily on clinical tests like electrocardiograms (ECGs), stress tests, and blood analysis. However, these methods are often reactive, detecting heart conditions only after they have progressed. In contrast, machine learning models like logistic regression can analyze large amounts of patient data to predict the likelihood of heart disease before symptoms manifest, offering a proactive approach to healthcare. With the increasing digitization of medical records and the growing availability of wearable devices, there is an unprecedented opportunity to leverage machine learning algorithms to predict heart disease risk at an early stage, enabling more personalized and preventive healthcare approaches.

II. Logistic Regression: An Overview:

Logistic regression is a supervised learning algorithm primarily used for binary classification. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that a given input belongs to a particular class. For heart disease detection, the two possible classes are typically “at risk” and “not at risk.”

The logistic function, or sigmoid function, maps predicted values to probabilities. This function constrains the output between 0 and 1, where values closer to 0 represent a lower probability of heart disease and values closer to 1 indicate a higher risk[3]. Logistic regression is highly interpretable and provides insights into the influence of each feature on the prediction. This makes it an excellent choice for medical applications, where understanding the contributing factors to a diagnosis is critical.

The model works by fitting a line to the data that best separates the two classes. It calculates the odds of a certain event (such as developing heart disease) happening, based on the input features. Each feature, such as age, cholesterol level, or smoking status, is assigned a weight that indicates its importance in predicting the outcome. These weights are optimized during the training process using techniques such as maximum likelihood estimation (MLE). Logistic regression is particularly favored in medical applications due to its simplicity, efficiency, and high interpretability, allowing healthcare professionals to understand which factors contribute most to a patient’s risk profile[4]. This transparency is crucial in fields like healthcare, where the interpretability of model decisions is often as important as the accuracy of the predictions.

III. Dataset and Feature Selection:

The accuracy and reliability of a machine learning model depend heavily on the quality of the dataset and the selected features. For heart disease prediction, datasets commonly include attributes such as age, gender, blood pressure, cholesterol levels, smoking habits, family history, and physical activity. These features are chosen because they are clinically relevant risk factors for heart disease[5]. Data preprocessing is a crucial step in ensuring the model's success. Missing values need to be imputed or removed, and features may require normalization or standardization to improve the performance of the logistic regression model. Feature selection techniques, such as recursive feature elimination or correlation matrices, can help identify the most predictive features, ensuring that the model remains both accurate and interpretable. The effectiveness of a logistic regression model in predicting heart disease

significantly hinges on the quality and relevance of the dataset and the selected features. In this study, we utilize several publicly available datasets, such as the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, which includes a variety of clinical and demographic information. Key attributes include age, gender, blood pressure, cholesterol levels, maximum heart rate achieved, and lifestyle factors like smoking status and physical activity. These features have been identified in medical literature as critical risk factors for heart disease.

Data preprocessing is vital for enhancing model accuracy and involves handling missing values, normalizing continuous variables, and encoding categorical variables[6]. For instance, the presence of null values in the dataset may be addressed through imputation methods, ensuring that the logistic regression model receives a complete dataset for training. Feature selection is another critical component, as it involves identifying the most predictive features that contribute to heart disease risk. Techniques such as correlation analysis, recursive feature elimination, or regularization methods (like LASSO) can help refine the feature set. By focusing on the most relevant features, we can enhance the model's performance, reduce overfitting, and improve interpretability, making it easier for healthcare practitioners to understand the underlying risk factors influencing heart disease predictions[7]. Ultimately, a well-curated dataset and thoughtful feature selection lay the foundation for developing a robust logistic regression model that can effectively identify patients at risk of heart disease.

IV. Model Training and Evaluation:

Training the logistic regression model involves finding the optimal parameters that minimize the loss function, typically using a method called maximum likelihood estimation (MLE). The model is trained on a labeled dataset, where the outcome variable indicates whether the patient has heart disease or not. The logistic regression algorithm iteratively adjusts its weights to better fit the data[8].

Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC-AUC) curve. These metrics give insight into how well the model is distinguishing between patients at risk of heart disease and those not at risk. Cross-validation techniques, like k-fold validation, can help ensure that the model generalizes well to unseen data.

Training a logistic regression model involves a systematic approach to optimizing its parameters for accurate predictions. The process begins by splitting the dataset into training and testing subsets, typically using an 80/20 ratio[9]. This ensures that the model can be evaluated on unseen data to gauge its generalizability. The training phase utilizes maximum likelihood estimation (MLE) to find the optimal coefficients for the model. During this phase, the algorithm iteratively adjusts these coefficients to minimize the discrepancy between the predicted probabilities and the actual outcomes. Once trained, the model's performance is assessed using various metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Each of these metrics provides insights into different aspects of model performance, such as its ability to correctly identify positive cases (sensitivity) and negative cases (specificity). Cross-validation techniques, such as k-fold cross-validation, can also be employed to ensure that the model remains robust across

different subsets of data, reducing the risk of overfitting. By analyzing these evaluation metrics, healthcare professionals can ascertain the model's reliability in predicting heart disease and make informed decisions on its integration into e-healthcare systems[10].

V. Comparison with Other Machine Learning Models:

While logistic regression is a simple and interpretable model, it is essential to compare its performance with other machine learning techniques such as decision trees, random forests, support vector machines (SVMs), and neural networks. Each of these models has its strengths and weaknesses. For instance, random forests and SVMs may provide higher accuracy, but they often lack the interpretability of logistic regression. In healthcare, model interpretability is critical, as medical professionals need to understand the rationale behind a prediction. Logistic regression strikes a balance between performance and explainability, making it a practical choice for heart disease prediction, especially when combined with other ensemble techniques for enhanced performance.

When evaluating the efficacy of logistic regression for heart disease prediction, it is essential to compare it with other machine learning models to understand its relative strengths and weaknesses. While logistic regression is simple and interpretable, more complex models, such as decision trees, random forests, and support vector machines (SVMs), can often yield higher accuracy[11]. Decision trees are intuitive and can handle nonlinear relationships, making them useful for understanding interactions between features. Random forests, an ensemble method that builds multiple decision trees, can mitigate overfitting and generally perform well on diverse datasets. SVMs, on the other hand, are effective for high-dimensional data but can be less interpretable than logistic regression. Neural networks have gained popularity for their ability to model intricate patterns in large datasets; however, they require significant amounts of data and computational power. This complexity can be a disadvantage in healthcare applications, where transparency is crucial for clinical decision-making. Logistic regression stands out because it offers a good balance between accuracy and interpretability, making it easier for healthcare professionals to understand the influence of various risk factors on heart disease predictions[12]. This interpretability is particularly important in medical contexts, where the rationale behind a prediction can influence patient management decisions. Ultimately, while other models may provide enhanced predictive performance, logistic regression remains a practical choice for early heart disease detection, especially in scenarios where explainability and ease of use are paramount.

VI. Applications in E-Healthcare:

The implementation of logistic regression models in e-healthcare systems can significantly impact how heart disease is detected and managed. Such models can be integrated into mobile health applications, wearable devices, or online platforms, enabling patients and healthcare providers to monitor heart health continuously[13]. For example, an app might prompt users to input daily data such as blood pressure and physical activity, generating a real-time prediction of heart disease risk. Telemedicine platforms can also benefit from these models by offering automated risk assessments, which could help doctors prioritize high-risk patients for further

diagnostic tests or interventions. E-healthcare systems that leverage logistic regression models can democratize access to high-quality healthcare, particularly in underserved areas.

The integration of logistic regression models in e-healthcare systems has transformative potential for early heart disease detection and management. By embedding these predictive models into mobile health applications and telemedicine platforms, healthcare providers can deliver real-time risk assessments to patients. For instance, users can input vital data such as blood pressure, cholesterol levels, and lifestyle factors, which the logistic regression model can analyze to generate immediate risk predictions[14]. This empowers patients to actively engage in their health management by encouraging them to adopt lifestyle changes or seek medical advice based on their assessed risk levels. Furthermore, healthcare providers can use these models to prioritize high-risk patients for follow-up appointments or diagnostic testing, thereby optimizing resource allocation and improving overall care efficiency. The scalability of e-healthcare solutions also enables broader access to preventative measures, particularly in rural or underserved areas, where traditional healthcare services may be limited. Overall, the application of logistic regression in e-healthcare not only enhances early detection but also fosters a proactive approach to heart health, ultimately leading to better patient outcomes and reduced healthcare costs.

VII. Challenges and Future Directions:

While logistic regression is a powerful tool, there are challenges associated with its use in heart disease prediction. One major challenge is the presence of imbalanced datasets, where the number of patients without heart disease far exceeds those with heart disease[15]. This imbalance can skew the model's predictions towards the majority class, reducing its ability to identify high-risk patients accurately. Techniques such as oversampling, undersampling, or synthetic data generation (e.g., SMOTE) can address this issue.

Moreover, the healthcare industry is characterized by constantly evolving data, such as the emergence of new biomarkers for heart disease. Future research could explore how logistic regression models can be continuously updated as new data becomes available, potentially through online learning techniques[16]. The integration of logistic regression with other machine learning models in ensemble frameworks also presents opportunities for improving predictive accuracy.

VIII. Conclusion:

Logistic regression offers a robust, interpretable, and efficient approach to early heart disease detection. Its ability to classify patients based on risk factors makes it highly suitable for e-healthcare applications, where timely intervention can save lives. While more complex models may offer improved accuracy, the interpretability of logistic regression ensures that it remains a valuable tool in healthcare, where understanding the basis of predictions is as crucial as the predictions themselves. Future research should focus on addressing challenges such as data imbalance and model updates, ensuring that logistic regression continues to evolve as a key tool in the fight against heart disease.

REFERENCES:

[1] S. S. Kute, A. Shreyas Madhav, S. Kumari, and S. Aswathy, "Machine learning–based disease diagnosis and prediction for E-healthcare system," *Advanced analytics and deep learning models,* pp. 127-147, 2022.

[2] M. R. Pulicharla and V. Premani, "AI-powered Neuroprosthetics for brain-computer interfaces (BCIs)," *World Journal of Advanced Engineering Technology and Sciences,* vol. 12, no. 1, pp. 109-115, 2024.

[3] S. Kumar, S. Srivastava, S. Mongia, and M. Amsa, "Diagnosis of heart disease using machine learning classification technique in e-healthcare," *Journal of Pharmaceutical Negative Results,* pp. 656-664, 2023.

[4] S. Khan and Z. Ali, "Deep Learning in Neuroprosthetics: Improving the Precision and Responsiveness of Brain-Machine Interfaces," *Innovative Computer Sciences Journal,* vol. 10, no. 1, 2024.

[5] V. Janarthanan, T. Annamalai, and M. Arumugam, "Enhancing healthcare in the digital era: A secure e-health system for heart disease prediction and cloud security," *Expert Systems with Applications,* vol. 255, p. 124479, 2024.

[6] F. Dahan, R. Alroobaea, W. Y. Alghamdi, M. K. Mohammed, F. Hajjej, and K. Raahemifar, "A smart IoMT based architecture for E-healthcare patient monitoring system using artificial intelligence algorithms," *Frontiers in Physiology,* vol. 14, p. 1125952, 2023.

[7] M. R. Pulicharla and A. Singhal, "Techniques for Machine Learning: Identifying Heart Disease within E-Healthcare through Implementation: Logistic Regression Model."

[8] M. Anshori and M. S. Haris, "Predicting heart disease using logistic regression," *Knowledge Engineering and Data Science (KEDS),* vol. 5, no. 2, pp. 188-196, 2022.

[9] M. A. Lebedev and M. A. Nicolelis, "Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation," *Physiological reviews,* vol. 97, no. 2, pp. 767-837, 2017.

[10] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE access,* vol. 8, pp. 107562-107582, 2020.

[11] S. Luo, Q. Rabbani, and N. E. Crone, "Brain-computer interface: applications to speech decoding and synthesis to augment communication," *Neurotherapeutics,* vol. 19, no. 1, pp. 263-273, 2023.

[12] M. Nasr, M. M. Islam, S. Shehata, F. Karray, and Y. Quintana, "Smart healthcare in the age of AI: recent advances, challenges, and future prospects," *IEEE Access,* vol. 9, pp. 145248-145270, 2021.

[13] R. M. Rothschild, "Neuroengineering tools/applications for bidirectional interfaces, brain–computer interfaces, and neuroprosthetic implants–a review of recent progress," *Frontiers in neuroengineering,* vol. 3, p. 112, 2010.

[14] M. Sliwowski, "Artificial intelligence for real-time decoding of motor commands from ECoG of disabled subjects for chronic brain computer interfacing," Université Grenoble Alpes [2020-....], 2022.

[15] S. R. Soekadar *et al.*, "Future developments in brain/neural–computer interface technology," in *Policy, identity, and neurotechnology: the neuroethics of brain-computer interfaces*: Springer, 2023, pp. 65-85.

[16] X. Zhang *et al.*, "The combination of brain-computer interfaces and artificial intelligence: applications and challenges," *Annals of translational medicine,* vol. 8, no. 11, 2020.