Comprehensive Study on Machine Learning Applications for Heart Disease Risk Prediction

Lashyn Adiat¹, Aizhan Altaibek¹², Marat Nurtas¹² and Aigerim Altayeva³

¹International Information Technology University, Manas St. 34/1, Almaty, 050000, Kazakhstan
²Institute of Ionosphere, Gardening community IONOSPHERE 117, Almaty, 050020, Kazakhstan
³Al-Farabi Kazakh National University, al-Farabi Avenue 71, Almaty, 050040, Kazakhstan

Abstract
This research investigates an analytical approach to machine learning applications for cardiovascular disease (CVD) risk prediction on a publicly accessible database. The dataset contains crucial information about patients’ individual characteristics, such as age, blood pressure, ECG at rest, heart rate, and four types of chest pain. The purpose of this research is to choose the most suitable model for heart attack analysis. Descriptive analytics and exploratory data analysis based on various factors were done to predict risk by employing machine learning algorithms and techniques, including k-nearest neighbours, logistic regression, support vector machines (SVM) and random forests. The research involves thorough data analysis and rigorous model training processes.

Keywords
Heart disease prediction, comprehensive study, machine learning, cardiovascular risks, exploratory data analysis

1. Introduction

The World Health Organization (WHO) classifies and reports on various causes of death worldwide through its International Classification of Diseases (ICD) system [1].

According to WHO, Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age [2].

Cardiovascular diseases (diseases of the heart or blood vessels) have become a significant public health concern in economically advanced countries. This is primarily due to the difficulties in making an early diagnosis and patients’ unwillingness to seek medical assistance when the first symptoms occur. A fast-paced lifestyle, an unhealthy diet, a lack of physical activity, alcohol and tobacco addictions, and insufficient sleep all contribute to the harmful influence on the cardiovascular system [3].

According to epidemiological data, in 2018, cardiovascular disease was the leading cause of death in China. The number of patients with cardiovascular disease is 330 million in China, including 11 million stroke and more than 270 million diseases related to the heart [4].

In the United States, heart disease causes more than 600,000 deaths annually, accounting for approximately one in every four deaths [5].

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms, and signs of the patient [6]. Body cholesterol levels, smoking habits, obesity, family history of diseases, blood pressure, and working environment are all factors that influence heart disease [10].
Early detection and prevention of heart attacks can improve patient outcomes and reduce the strain on healthcare systems.

Traditional risk assessment methods, such as the Framingham Risk Score [7], focus on clinical and demographic data that may not fully capture the complexity of the underlying causes of cardiovascular risk.

Machine learning has transformed disease detection by enabling the creation of predictive models [3] that analyze massive datasets to uncover subtle trends and anomalies, thereby assisting in early diagnosis and intervention. Machine learning techniques have been extensively used in recent years to forecast the likelihood of heart attacks based on these parameters [8].

Although traditional risk assessment models are effective, their scope and predictive power are frequently constrained. In contrast, machine learning offers a data-driven method that can more accurately forecast cardiac disease and find complex correlations between numerous risk factors.

As healthcare organizations strive to acquire patient records, it is estimated that one trillion bytes of data are generated every day. This information is of the utmost importance and must be properly extracted to yield valuable insights [13]. Patients may not always accurately describe their medical conditions, and laboratory test results can be subject to errors. Healthcare specialists may struggle to make informed decisions about a patient's illness because of their limited expertise in all areas [12]. To address this challenge, the development of a disease prediction system that integrates medical knowledge with a comprehensive system is necessary to produce the most effective results and benefit society [14]. Previous investigations have attempted to use patient laboratory tests [15-17] and medication [18] to predict disease onset. Some prototypes have also been used to identify unknown risk factors while simultaneously improving the sensitivity and specificity of detection. Recent studies have demonstrated success in predicting diseases through several methods, including support vector machines [19-21], logistic regression [22], random forests [23], neural networks [17], and time series modelling techniques [24].

Machine learning models can be flexible and tailored to fit the range of data sources that are becoming increasingly accessible in healthcare. The advancement in technology has greatly enhanced the capacity to forecast a wide range of diseases, such as cancer, cardiovascular conditions, and infectious outbreaks. This development has a two-fold impact: it empowers healthcare providers to identify high-risk individuals for early intervention, which could potentially save lives, and it supports public health agencies in proactive surveillance and resource allocation, helping to curb the spread of diseases on a larger scale. As machine learning continues to advance, its contribution to disease prediction is expected to improve healthcare outcomes and minimize the avoidable economic and human costs associated with preventable illnesses [9].

The goal of this research is to build and test a machine-learning model for forecasting the risk of heart disease. This will be accomplished using a dataset containing patient information and clinical measurements. A comparative analysis is performed to evaluate and contrast the efficacy of various machine learning algorithms that have been utilized in this context, including logistic regression and more advanced models and feature selection strategies in the domain of heart attack prediction. The selection of the most appropriate algorithm is of the utmost importance, as it has a direct impact on the model's ability to process the data, recognize complex relationships, and produce trustworthy predictions.

Considering the study's findings, it is critical to provide relevant insights and evidence-based recommendations to healthcare providers and policymakers. The scope of the investigation is limited to the assessment of a single dataset containing patient information and clinical measurements. The model developed in this study was built primarily for research purposes and should not be used in place of clinical diagnosis or therapy.

2. Methodology
The exploratory data analysis was performed using publicly accessible data on heart disease. The dataset comprised 303 records with 14 attributes, including age, blood pressure, blood glucose level, ECG at rest, heart rate, and four types of chest pain.

![Table 1](#)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age of the patient</td>
</tr>
<tr>
<td>sex</td>
<td>Sex of the patient</td>
</tr>
<tr>
<td>cp</td>
<td>Four chest pain types:&lt;br&gt;0 = Typical Angina&lt;br&gt;1 = Atypical Angina&lt;br&gt;2 = Non-anginal Pain&lt;br&gt;3 = Asymptomatic</td>
</tr>
<tr>
<td>trtbps</td>
<td>Resting blood pressure (in mm Hg)</td>
</tr>
<tr>
<td>chol</td>
<td>Cholestral in mg/dl fetched via BMI sensor</td>
</tr>
<tr>
<td>fbs</td>
<td>(fasting blood sugar &gt; 120 mg/dl) ~ 1 = True, 0 = False</td>
</tr>
<tr>
<td>restecg</td>
<td>Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy</td>
</tr>
<tr>
<td>thalachh</td>
<td>Maximum heart rate achieved</td>
</tr>
<tr>
<td>exng</td>
<td>Exercise induced angina ~ 1 = Yes, 0 = No</td>
</tr>
<tr>
<td>oldpeak</td>
<td>Previous peak</td>
</tr>
<tr>
<td>slp</td>
<td>Slope</td>
</tr>
<tr>
<td>caa</td>
<td>Number of major vessels</td>
</tr>
<tr>
<td>thall</td>
<td>Thalium Stress Test result ~ (0,3)</td>
</tr>
<tr>
<td>output</td>
<td>Target variable</td>
</tr>
</tbody>
</table>

The nearly balanced data on the proportion of people experiencing heart attacks (54%) suggests that there is no need to further balance them.
The majority of individuals fall within the age range of 50-60 years old, have relatively low chest pain, have blood pressure within the range of 120-140, have cholesterol levels between 200-300, have blood sugar levels below 120, and are male. The majority of these individuals had a heart rate within the range of 150-175.

Individuals aged 40-60 are more likely to have heart disease, whereas those with a higher resting heart rate are at a higher risk of experiencing a heart attack.
Figure 4: Heart Attack distribution over age and heart rate

People having cholesterol of 120-250 and blood between 110 to 140 are more likely to have a heart attack.

Figure 5: Heart Attack distribution over cholesterol and blood pressure

A large percentage of men are more likely to experience heart attacks than women, with 73% of men and 45% of women suffering from heart attacks.
If someone experiences chest pain, it is highly probable that they will suffer from a heart attack.

The impact of blood sugar level on the likelihood of a heart attack is relatively small. In other words, whether a person has high blood sugar levels does not necessarily determine whether they will have a heart attack.

People who do not regularly exercise their cardiovascular system are highly likely to suffer from heart attack.

Figure 6: Distribution of features with reference to Heart Disease

The higher the chest pain and the higher the person's heart rate, the more likely they are to suffer a heart attack.

Figure 7: Distribution of attribute cp with reference to other attribute thalachh having hue=Heart Attack

Individuals with a low level of exercise-induced angina are more likely to experience heart disease, even though age does not significantly contribute to the risk of heart attack.
The graph and table below show that there is a positive correlation between heart attack and chest pain, heart rate, and slope. However, there was a negative correlation between heart attacks and age, induced angina, and major vessels.
3. Results

Among the aforementioned machine learning algorithms, the utilization of logistic regression, K-nearest neighbours (KNN) [11], support vector classifiers, and random forest classifiers are recommended because they demonstrate comparative accuracy with traditional methods. Our preliminary examination suggests that the search for optimal coefficients may be accommodated through iterative coefficient selection, specifically, Logistic Regression. However, the accuracy of the models can only be ascertained after a thorough evaluation.

![Figure 10: Diagram of accuracies by machine learning methods](image)

Logistic Regression, a statistical method used for binary classification, was first introduced to address the problem of binary classification. It assumes that the data follow a Bernoulli distribution and solves for the optimal parameters through maximum likelihood estimation [4]. The first Logistic Regression model shows values of around 89 per cent accuracy, which is quite good for this model.

![Figure 11: Roc Curve for logistic regression](image)

KNN demonstrates an accuracy of around 70%, which unfortunately means that finding the optimal K requires additional calculations up to K=100(th neighbour). SVC is showing progress and is approaching 76% but it is still not the best model at the moment. Finally, Random Forest shows progress around 82%, which is clearly better than the previous ones, but still falls short of the best model.

We can conclude that the Logistic Regression model shows good progress in comparison with other machine learning algorithms, thereby suggesting that the neural network will demonstrate a more optimal result since Logistic Regression is based on gradient descent in finding the necessary parameters.
4. Acknowledgements

We would like to express our sincere gratitude to Rashik Rahman, the data provider on Kaggle for providing valuable data resources that were essential for the successful completion of this study. We extend our appreciation to the Kaggle community for their efforts to make diverse datasets accessible and to promote collaboration among data enthusiasts and researchers.

5. Conclusion

During our study, we conducted an exploratory data analysis and comparative analysis of machine learning algorithms’ accuracies. After conducting a thorough analysis and applying several widely used machine learning algorithms for forecasting heart disease, we discovered that logistic regression demonstrated outstanding performance. In fact, it was the algorithm that attained the highest level of accuracy, allowing us to confidently classify patients with heart disease. The application of machine learning algorithms in predicting heart disease is an ongoing research area with significant promise. Integrating advanced machine learning methods in this domain is likely to significantly alleviate the burden on health care and improve the prognosis of diseases, leading to improved overall patient health.

Therefore, in future research, we aim to develop personalized forecasting methods that consider additional risk factors and model adaptations to changing conditions and patient needs. This approach will enhance the accuracy and effectiveness of heart disease forecasts, leading to earlier diagnosis and treatment, and ultimately improving patient health and quality of life.

6. References


