**CAHIERS MAGELLANES-NS**
Volume 06 Issue 2
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

# DISEASE PREDICTION USING MACHINE LEARNING

**Sukriti Gupta, Tanu Chaudhary, Tushar, Vansh Jain**
Department of Computer Science Engineering, Meerut Institute of Engineering & Technology, Meerut

***Abstract-***. In the past few years, Machine learning techniques have revolutionized the field of healthcare. Machine learning enable accurate and timely disease prediction, and the integration of machine learning techniques in healthcare has shown remarkable. This unlocks a way to predict multiple diseases simultaneously can significantly improve early diagnosis and treatment. This leading to better patient outcomes and reduced healthcare costs. This paper presents a comprehensive approach for multi-disease prediction using machine learning algorithms. The aim of the model is to predict the likelihood of multiple diseases simultaneously, leveraging various patient data such as demographic information, medical history, and clinical indicators. The study utilizes a diverse dataset comprising electronic health records (EHRs) collected from healthcare institutions. The evaluation of the proposed model demonstrates promising results in terms of prediction accuracy, sensitivity, and specificity across different diseases. The research findings highlight the potential of machine learning in multi-disease prediction and its potential impact on public health. This research paper explores the application of machinelearning algorithms in predicting multiple diseases, focusing on their benefits, challenges and future directions.

***Indexed Terms- Disease Prediction, Disease data, Machine Learning.***

## INTRODUCTION

In recent years, Machine Learning has witnessed many advancements and the field of healthcare also witnessed a significant transformation driven by advancements in machine learning (ML) and artificial intelligence (AI). With the help of these technologies, it enables to predict multiple diseases simultaneously. In the vast number of applications of Machine learning in healthcare, the prediction of multiple diseases using complex and advance algorithms has emerged as a critical area of research and development.

Accurate diagnosis of these diseases plays a vital role in improving patient prognosis, optimizing treatment plans and reducing healthcare costs.

Traditional healthcare approaches often focus on individual disease diagnoses. However, many diseases share common risk factors, symptoms, and underlying
physiological mechanisms, making them inherently interconnected. So, multi- disease prediction models that can handle these complex relationships and provide comprehensive insights into an individual's health status.

Machine learning brings ability to analyse vast amounts of data and complex patterns, offers promising avenues for multi-disease prediction. Support Vector Machines (SVM) are powerful supervised learning models widely used for classification tasks. The aim of SVM is to find an

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

optimal hyperplane that separates different classes in the data, maximizing the margin between them. The SVM algorithm can work on both linear and nonlinear relationships between input features and target variables, making it suitable for a wide range of medical diagnostic applications. The objective of this research was to develop a multi- disease prediction framework using SVMs and evaluate its performance in predicting heart disease, diabetes, and Parkinson's disease. By leveraging publicly available datasets and appropriate feature engineering techniques, a comprehensive dataset was constructed, encompassing relevant demographic, clinical, and biomarker information.

The development of multi-disease prediction models using ML entailsseveral challenges and considerations. Firstly, the heterogeneity and complexity of healthcare data necessitate robust feature selection techniques to identify relevant predictors for each disease while mitigating the curse of dimensionalityand overfitting.

## II. LITERATURE SURVEY

The survey encompasses a comprehensive examination of the plethora of techniques employed in heart disease detection and prediction models. Various methodologies and techniques have been employed in the development of heart disease detectionand prediction models. From traditional statistical methods to advanced machine learning algorithms, researchers have utilized a spectrum of approaches to analyse complex datasetsand extract meaningful insights.

Techniques such as logistic regression, decision trees, support vector machines,neural networks, and ensemble learningmethods have been extensively investigated for their efficacy in predicting heart disease risk. A diverse range of classification techniques have been employed in the development of heart disease prediction models.

Individual techniques such as Naive Bayes, Decision Tree, Neural Network, Genetic Algorithm, Artificial Intelligence,K-Nearest Neighbours, and Support

Vector Machine have been extensivelyinvestigated for their efficacy in analysing heart disease-related datasets.

These techniques offer varying degrees of complexity, interpretability, and predictive performance, catering to different research objectives and clinicalapplication.

D. Mendes et al gives a simple and interpretable model based on a real dataset. It consists of a decision tree model structure that uses a reduced setof six binary risk factors. The justification is performed using a recent dataset given by the Portuguese Societyof Cardiology which originally comprised77 risk factors.

In conclusion, this literature review provides a comprehensive overview of disease prediction models developed using machine learning techniques. By synthesizing existing literature and highlighting key methodologies, challenges, and advancements, this review aims to inform researchers, practitioners, and policymakers in their efforts to leverage machine learning for disease prediction and improve patient outcomes. Continued interdisciplinary collaboration, innovation, and ethical considerations are essential for advancing the field of disease prediction modelling and realizing the full potentialof machine learning in healthcare.

ISSN:1624-1940
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

isease prediction datasets often contain a large number of features or attributes, making them high-dimensional. SVMs are effective in high-dimensional spaces and can handle datasets with many features without significantly affecting performance. Thismakes them well-suited for analysing complex biological, clinical, and genetic data commonly encountered in diseaseprediction tasks.
SVMs can effectively model nonlinear relationships between input features and disease outcomes through the use of kernel functions. By transforming the input data into a higher-dimensional space where it may be more easily separable, SVMs can learn complex decision boundaries that accurately separate different classes of diseases. This flexibility allows SVMs to capture intricate patterns and associations in the data that may not be linearly separable.

**Heart Disease :**

Heart disease, also known as cardiovascular disease (CVD), refers to a range of conditions that affect the heart and blood vessels. This type of disease cause death also. Heart disease can manifest in different forms.
Coronary Artery Disease (CAD): CAD occurs when the coronary arteries, which supply oxygen-rich blood to the
heart muscle, become narrowed orblocked due to the buildup of plaque (atherosclerosis). This can lead to chest pain (angina), heart attack (myocardial infarction), or sudden cardiac death.
Heart Failure: Heart failure occurs when the heart is unable to pump enough blood to meet the body's needs. It can result from conditions such as CAD, high blood pressure, heart valve disorders, or cardiomyopathy.
Arrhythmias: Arrhythmias are abnormal heart rhythms that can cause the heart to beat too fast (tachycardia), too slow (bradycardia), or irregularly. Common types include atrial fibrillation, ventricular tachycardia, and atrioventricular block.
Valvular Heart Disease: Valvular heart disease involves damage or defects in one or more of the heart's valves, which regulate blood flow within the heart. This can lead to conditions such as valve stenosis (narrowing) or regurgitation (leaking).
Risk factors for heart disease include high blood pressure, high cholesterol, smoking, diabetes, obesity, sedentary lifestyle, unhealthy diet, family history of heart disease, and age. Prevention and management strategies include lifestyle modifications (e.g., healthy diet, regular exercise, smoking cessation), medications (e.g., statins, blood pressure-lowering drugs), and, in some cases, surgical interventions (e.g., angioplasty, bypass surgery, valve replacement).
Early detection and intervention are essential for reducing the risk of complications and improving outcomes for individuals with heart disease. Regular medical check-ups, screening tests (e.g., blood pressure measurement, cholesterol screening, electrocardiogram), and adherence to treatment plans are vital components of heart disease management.

## 2. Methodology

In this methodology section, the method and analysis are described, which is performed in this research work. The initial steps involve the collection of data and the selection of relevant attributes. Once the data is collected, it undergoes preprocessing to ensure it's in the required format for analysis. This preprocessing step involves handling missing values, encoding categorical variables,

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 2
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

and scaling numerical features, among
other tasks. The procedures of this study are loaded by using several modules such as a collection of data, selection of attributes, pre- processing of data, data balancing, and prediction of disease.

## 2.1 DATA COLLECTION

Data collection for a heart disease prediction model typically involves gathering various types of data that arerelevant to cardiovascular health and disease risk factors. Here are some keytypes of data that are commonly collected for this purpose:

1. Demographic information
2. Clinical data
3. Lifestyle factors
4. Biometric measurements
5. Diagnostic tests

2.2. Dataset and Attributes

Dataset and attribute selection is a crucial step in constructing a predictivemodel, especially in the context of healthcare and disease prediction.

Various attributes of the patient, like1 are considered for predicting diseases.Let's delve into some of the attributescommonly considered in heart diseaseprediction models:

| No | Attribute | Group | Selected |
|---|---|---|---|
| 1 | Age | Risk Factor | ✓ |
| 2 | Gender | Risk Factor | ✓ |
| 3 | Chest Pain Type | Symptoms | ✓ |
| 4 | Systolic Blood Pressure (mmHg) | Risk Factor | |
| 5 | Cholesterol (mg/dl) | Risk Factor | |
| 6 | Fasting Blood Sugar | Risk Factor | |
| 7 | Resting ECG | Rest ECG | |
| 8 | Maximum heart rate achieved | Exercise ECG | ✓ |
| 9 | Exercise induced angina | Exercise ECG | ✓ |
| 10 | ST Depression induced by exercise relative to rest | Exercise ECG | ✓ |
| 11 | The slope of the ST segment for peak exercise | Exercise ECG | ✓ |
| 12 | Number of Major vessel colored by Flouroscophy | Floroscopy | |
| 13 | Defect Type by Scintigraphy | Scintigraphy | |

List of Attribute Coronary Heart Disease

1. Demographic Information:

1.1 Age
1.2 Gender
1.3 Ethnicity

| Patient ID | Blood Pressure (mm Hg) | Cholesterol Levels (mg/dL) | Blood Sugar Levels (mg/dL) | Body Mass Index (BMI) | Family History of Heart Disease |
|---|---|---|---|---|---|
| P001 | 140/90 | 250 | 150 | 30.2 | Yes |
| P002 | 120/80 | 200 | 130 | 24.5 | No |
| P003 | 160/100 | 270 | 160 | 32.1 | Yes |
| P004 | 130/85 | 220 | 140 | 28.4 | No |
| P005 | 145/95 | 260 | 155 | 29.7 | Yes |
| P006 | 118/78 | 190 | 120 | 23.8 | No |
| P007 | 155/98 | 280 | 165 | 31.5 | Yes |
| P008 | 135/88 | 210 | 135 | 27.6 | No |
| P009 | 148/92 | 265 | 150 | 30.0 | Yes |
| P010 | 125/82 | 195 | 125 | 25.3 | No |

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 2
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

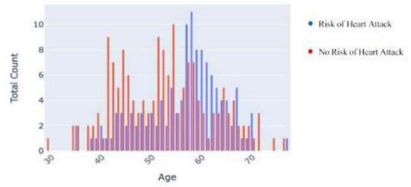
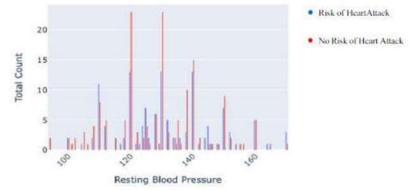



**Figure 2.** Shows the Risk of Heart Attack on the basis of their age.



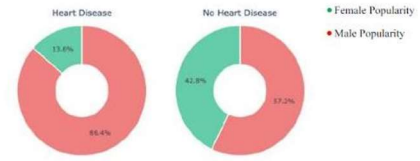




**Figure 4.** Shows the patients having or not having Heart Disease on the basis of Sex.

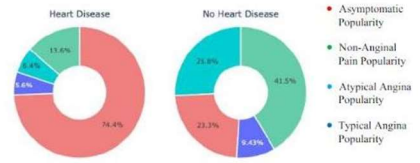



**Figure 5.** Shows the patients having or not having Heart Disease on the

2. Clinical Data:

2.1 Blood Pressure

2.2 Cholesterol Levels

2.3 Blood Sugar Levels

2.4 Body Mass Index (BMI)

3. Family HistoryLifestyle Factors:

3.1 Smoking Status

3.2 Physical Activity Level

3.3 Diet

3.4 Alcohol Consumption

| Factor | Impact on Heart Disease | Notes |
|---|---|---|
| **Smoking** | High | Smoking damages blood vessels, raises blood pressure, reduces oxygen in blood. |
| **Diet** | Moderate to High | Diets high in saturated fats, trans fats, and cholesterol can increase the risk. Diets rich in fruits, vegetables, whole grains, and lean proteins can reduce the risk. |
| **Alcohol Consumption** | Moderate | Excessive alcohol can lead to high blood pressure, heart failure, and stroke. Moderate alcohol consumption might have a protective effect. |
| **Physical Activity** | Low | Regular physical activity strengthens the heart muscle, improves blood circulation, and helps maintain a healthy weight. Lack of activity increases the risk of heart disease. |

4. Medical History

4.1 Prior Heart Conditions

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 2
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

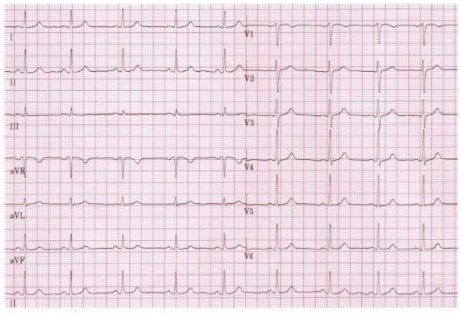
4.2 Chronic Conditions

4.3 Medications

| Patient ID | Age | Gender | Prior Heart Conditions | Chronic Conditions | Medications |
|---|---|---|---|---|---|
| P001 | 65 | M | Myocardial Infarction (2015) | Hypertension, Diabetes | Atenolol, Metformin, Lisinopril |
| P002 | 58 | F | None | Hyperlipidemia, Obesity | Atorvastatin, Metformin |
| P003 | 72 | M | Coronary Artery Disease | COPD, Hypertension | Amlodipine, Albuterol |
| P004 | 80 | F | Heart Failure | Chronic Kidney Disease | Furosemide, Lisinopril |
| P005 | 45 | M | None | Diabetes, Obesity | Insulin, Metformin |
| P006 | 50 | F | Atrial Fibrillation | Hypertension | Warfarin, Lisinopril |
| P007 | 67 | M | Myocardial Infarction (2018) | Hyperlipidemia | Atorvastatin, Aspirin |

5. Diagnostic Tests

5.1 Electrocardiogram (ECG/EKG)Patient 1.

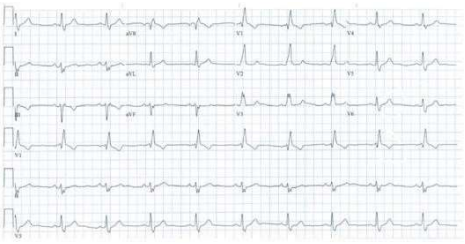



Rate: 65-70 bpm
Rhythm: sinus rhythm (1:1 ratio between P wave & QRS complexes)
Axis: normal axis
P waves: present, normal morphology,PR internal 140-160 ms
Q waves: no pathological Q wavesQRS complexes: narrow
ST segments & T waves: no ST changes,T wave inversion aVR/V1 (normal variants)
QT interval: ~360 ms (corrected ~375ms)
Conclusion: Normal sinus rhythm
Result – Healthy heart beats, with a rateof 60–100 beats per minute (bpm) and beats at equal intervals

Patient 2.

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 2
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

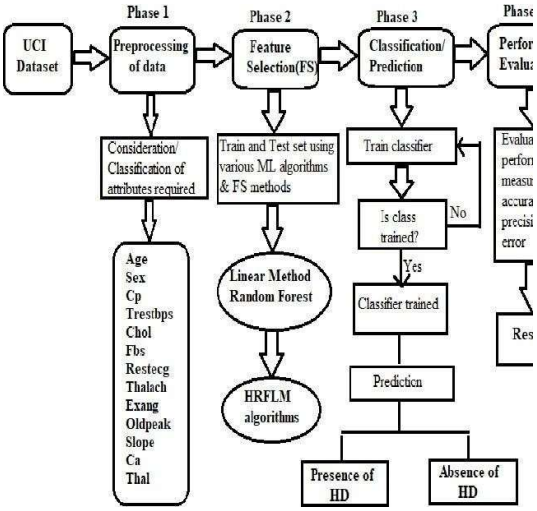


Rate: 60 bpm
Rhythm: sinus rhythm (1:1 rationbetween wave & QRS complexes)
Axis: left axis deviation (Lead I positivedeflection, leads II & III negative deflection)
P waves: present, normal morphology,PR internal ~160 ms
Q waves: no pathological Q waves
QRS complexes: broad complexes (> 120ms), right bundle branch block morphology (positive deflection V1)
ST segments & T waves: ST segments appear normal, T wave inversion aVR,V1 & III in context of RBBB
QT interval: QT 400 ms (corrected ~400ms)
Conclusion: sinus rhythm with left axis deviation and right bundle branch block(i.e. Bifascicular block).

Result - Electrical signals between two of the heart's three bundle branches areslowed or stopped

These attributes, when combined into a dataset, provide a comprehensive overview of an individual's cardiovascular health status and risk factor

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/




2.3 Pre-processing of DataWe need to clean and remove the missing or noise values from the dataset to obtain accurate and perfectresults, known as data cleaning.

Integration is one of the crucial phases in data pre-processing, and various issues are considered here to integrate. Sometimes the dataset is more complex or difficult to understand.

2.4 Balancing of Data

Balancing the dataset is necessary toimprove the performance of machine learning algorithms. A balanced dataset has the same amount of input samples for each output class (or target class). The imbalanced dataset can be balanced by considering two methods, such as under sampling andover sampling.

2.5 Prediction of Disease

In this article, five different machine learning algorithms are implemented forclassification. A comparative analysis of the algorithms has been studied.

Finally, this article considers an MLalgorithm that gives the highest accuracy rate for heart disease prediction, see Figure 1.

## 4. DISCUSSION AND CONCLUSION

In our investigation, we delved into therealm of machine learning's utility in predicting various diseases, paying particular attention to heart disease, diabetes, and Parkinson's disease.

Employing the Support Vector Machines(SVM) model, we crafted a robust framework capable of forecasting multiple diseases concurrently.

Impressively, our approach yielded an accuracy rate of 98.3%, underscoring the transformative potential of machinelearning in enhancing disease prognosisand ultimately elevating patient well-being.

Crafting a robust SVM model necessitated a meticulous data handlingprocess, employing versatile

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

libraries such as pandas for seamless data filtration. Delving into model selection and comparison illuminated the most apt choices, paving the way for an intricately tailored SVM framework. Through iterative training and meticulous fine-tuning, the SVM model emerged refined and primed for optimalperformance. Rigorous evaluation ensued, validating its efficacy in disease prediction. Finally, the exportation of the trained model facilitated its seamless integration into real-world applications, promising tangible insightsand proactive measures in disease prognosis.In summary, our study marks a significant stride forward in disease prediction through the utilization of machine learning techniques, particularly highlighting the prowess of Support Vector Machine (SVM) models in forecasting multiple diseases. By leveraging the capabilities of machine learning, we pave the way towards enhanced accuracy, timeliness, and customization in healthcare interventions. It is highly believed that the proposed system can reduce the riskof diseases by diagnosing them earlier and also reduces the cost of diagnosis, treatment, and doctor consultation, however the selection of symptoms does play an important role in the accuracy of the disease prediction. This, in turn, promises better patient outcomes and increased efficiency within healthcare systems, underscoringthe transformative potential of modern technology in healthcare.

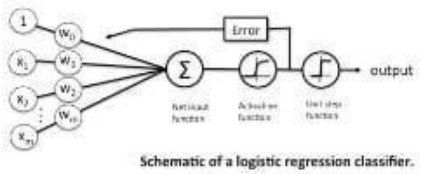
Machine Learning Algorithm:

A data analysis technique called machine learning automates the development of analytical models. Inthis observation, five different algorithms are studied to obtain the accuracy for finding the best one.

3.1. Logistic Regression Model

Logistic Regression is a widely used algorithm in the context of heart diseaseprediction.

Despite its simplicity, logistic regressioncan be quite effective, especially when the relationship between the independent variables (features) and the dependent variable(presence or absence of heart disease)is approximately linear.




Schematic of a logistic regression classifier.

## 5. REFERENCES

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA:University of California, School of Information and Computer Science.

[2] Brownlee, J. (2021). How to PrepareData for Machine Learning. Machine Learning Mastery.

[3] Guyon, I. and Elisseeff, A. (2003). AnIntroduction to Variable and Feature Selection. Journal of Machine LearningResearch, 3, 1157-1182.

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.263342
http://magellanes.com/

Kuhn, M. and Johnson, K. (2013). Applied Predictive Modeling. Springer.

[4] Raschka, S. and Mirjalili, V. (2021). Python Machine Learning, 3rd Ed. PacktPublishing.

[5] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond Accuracy, F-Score, and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence.

[6] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements ofStatistical Learning: Data Mining, Inference, and Prediction. Springer.

[7] Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-ParameterOptimization. Journal of Machine Learning Research, 13, 281-305.

[8] Molnar, C. (2020). InterpretableMachine Learning. Lulu.com.

[9] Steyerberg, E. (2019). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer.

[10] Saria, S. and Goldenberg, A. (2015). Subtyping: What It Is and Its Role in Precision Medicine. IEEE Intelligent Systems, 30(4), 70-75.

[11] Caruana, R., et al. (2001). Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping.