A PROJECT REPORT

on

"Heart Disease Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

ANKIT ANSHUMAN MOHAPATRA 2006006

ABHIJEET KUMAR 2006393

RAJDEEP KUNDU 2006514

ASHUTOSH YASH 2006518

UNDER THE GUIDANCE OF

MANAS RANJAN NAYAK

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

May 2020

KIIT Deemed to be University

School of Computer Engineering

CERTIFICATE

This is certify that the project entitled

"Heart Disease Prediction"

submitted by

Ankit Anshuman Mohapatra 2006006

Abhijeet Kumar 2006393

Rajdeep Kundu 2006514

Ashutosh Yash 2006518

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: 01 / 05 / 23

MANAS RANJAN NAYAK

Project Guide

Acknowledgements

We are profoundly grateful to **MANAS RANJAN NAYAK** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

Ankit Anshuman Mohapatra 2006006
Abhijeet Kumar 2006393
Rajdeep Kundu 2006514
Ashutosh Yash 2006518

CONTENT-

ABSTRACT

Over 17.9 million people die annually due to Cardiovascular Diseases(CVDs), representing 32% of all global deaths. Of these deaths, 85% are directly caused by heart attack and stroke. Earlier Machine learning algorithms were used to detect whether a person is suffering from CVDs by considering certain attributes like chest_pain, chol_level, age of the person & resting blood pressure levels. Previous researches were focused on identifying the significant features to heart disease prediction, however, less significance was given to contributing ones, which would be vital in identifying, visualizing, fitting and predicting the outcome to these features.

This paper emphasizes on the development of an AI-based heart disease diagnosis system using deep learning & machine learning algorithms. This study measures the strength of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, which is aimed at predicting CVDs based on the scores of the contributing features. Here in this research I predicted the scores of different algorithms contributing to heart disease prediction using the 300+ sample & 14 attribute, The University of California Irvine Heart Disease Data set. The diagnostic application achieves highest confidence score/accuracy of 90.16% in detecting heart disease. This research has managed to provide a leap through in predicting heart disease using machine learning & artificial intelligence.

Keywords:
Heart Disease Prediction, Cardiovascular Diseases (CVDs), Machine Learning Algorithms (KNN, SVM, Random Forest & Decision Tree), Deep Learning (Artificial Neural Network), The University of California Irvine Dataset.

School of Computer Engineering, KIIT, BBSR

SECTION - 1 Introduction

Cardiovascular Diseases(CVDs) are often referred to as heart disease. These diseases often refer to clogged or narrowed arteries that could lead to stroke, chest pain or angina, and other CVDs. Other types of heart disease, such as those that affects the heartbeat, valves, or muscles, are other types of heart disease. Machine learning, on the other hand, is important in determining whether a person has heart disease. But if these are anticipated, it will be easier for doctors to access crucial and vital information for treatment and diagnosing of patients. Heart disease is often a false symptom of coronary artery disease and hence, can be fatal.

Python is a content-oriented programming language that is well-designed and can be developed quickly. According to my analysis it considered one of the best programming language and has many applications in medicine. It is also considered a popular and well-received language with applications driven by AI-based software development and many other web applications. As mentioned the python framework is easy to use for building desktop or web-based applications. According to the use of python programming in medicine, especially in cardiac diagnosis, doctors and hospitals can better benefit patients by using a more capable, scalable and dynamic system. However, the coding packages and libraries used in this project are Pandas, matplotlib, Numpy, plot.py, seaborn and other sklearn features etc.

Data mining provides many techniques to discover hidden patterns or s imilarities in data. Therefore, in this research paper, several ML algorithms are proposed, which will be used to a validate heart disease prediction system on two open-access disease datasets.

SECTION - 2
Basic Concepts/ Literature Review

This project involves detecting heart disease using Python via ML & AI. The data set include 14 attributes. Project, matplotlib, Numpy, Pandas, alert etc. It uses several other libraries such as Correlation matrices, histograms, support vector classifiers, K-neighbor classifiers, random forest classifiers, and decision tree classifiers are used to evaluate the results of the data set using the python language. In addition to that, Python is considered an open source programming language, which supports the development of new solutions for healthcare and provides better outcomes for patients, thereby improving health.

One of the best-known ML algorithm tasks is data pre-preprocessing and classification. In this context, ML is often important for extracting information from involved businesses and transferring it to larger datasets. Most machine learning methods rely on these features that directly or indirectly affect the complexity of the model to explain the behavior of the algorithm.

The model will also be deployed in this project. After the data is loaded and saved in the variable used to copy the data. Finally, the dataset will be exported to file and processed. However, when the results are examined, it is seen that ANN algorithm scores 90.16%, while the support vector, Logistic Regression and random forest classifiers score 86% each, on the other hand decision tree gives 81.6%.

SECTION - 3
Problem Statement / Requirement Specification

Over 17.9 million people and countless others die annually, due to CVDs. So it is the need to build an efficient Machine Learning model, having the maximum accuracy to predict whether a person is diagnosed with heart disease or not! So in this section this SRS document gives the model planning, purpose, analysis & design that we have built.

3.1 Model Planning

The formulation of the model begins with gathering data from past research papers & creating a improved, idealistic & reliable model. Then we fulfilled the crucial requirement by selecting The University of California Irvine Heart Disease Data set & further optimized it.

3.2 Model Analysis

After the requirements are collected then it is loaded for Exploratory Data Analysis, under which we further normalized / standardized the data. Then using various models & techniques we optimized the models to get best accuracy.

3.3 Data Visualization

We Further Organized the data to perform Visualizations depicting the relationship between the attributes. Visualization gives a visual summary of the data set and customized plots such as Scatter plot, Missing number Matrix & Bar graphs are to embed into the applications, Here library's like Seaborn & Matplotlib is primarily used for statistical calculations on plots.

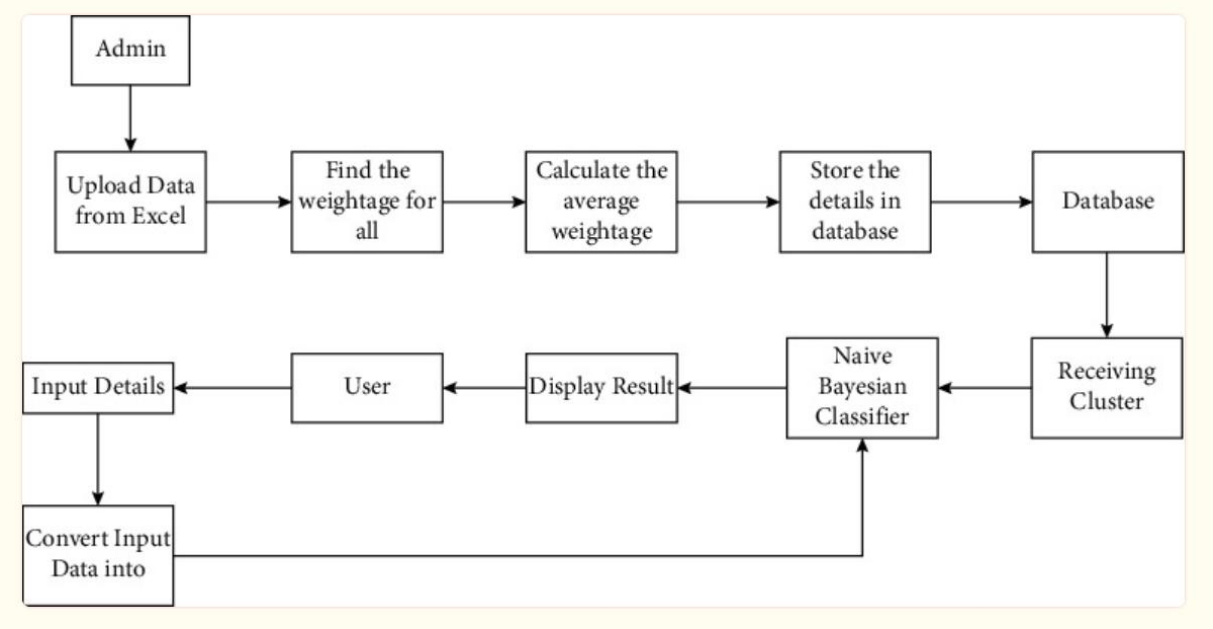
3.3.1 Design Constraints

Accuracy: The first constraint is that the visualization should accurately represent the data and the machine learning model's predictions. Any inaccuracies or distortions in the visualization can lead to incorrect conclusions and potentially harmful decisions, so we have managed through careful analysis.

Security: The second constraint is that the visualization should be designed with security in mind. Any sensitive information here is protected and encrypted to prevent unauthorized access or data breaches.

Responsiveness: The third constraint is that the visualization should be responsive to different devices and screen sizes. Here we have designed the model to adapt to different resolutions and orientations, ensuring that users can view and interact with the visualization on a variety of devices, including desktop computers, tablets, and smartphones.

3.3.1 System Architecture (UML Diagram)



SECTION - 4 Implementation

In this section, we have presented the implementation done by us during the model fitting & deployment.

4.1 Methodology / Procedure

This section explains the fitting of models and shows all the components, methods and tools used to in its development. Creating an intelligent and user-friendly heart disease predictor requires efficient tools to train large datasets and compare different machine learning algorithms. After choosing the most robust ANN algorithm which gives the maximum efficiency of 90.16, Which we will deploy in a web application using streamlit via the spyder IDE.

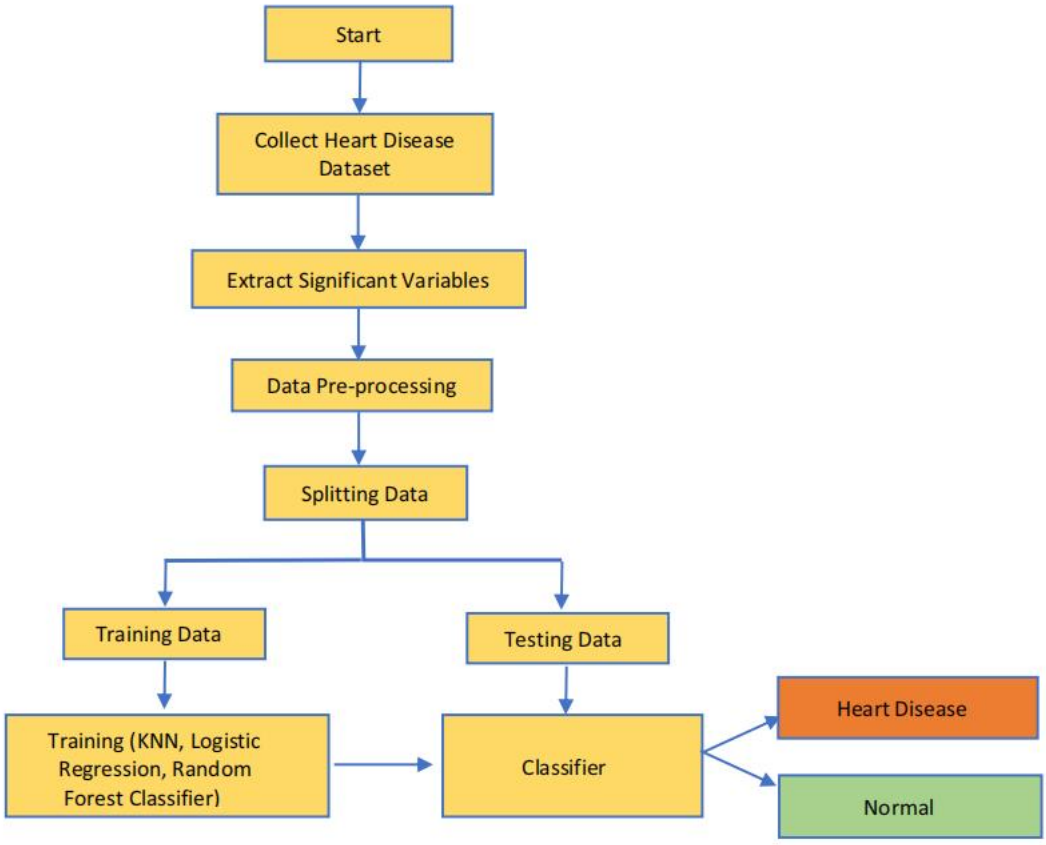


Figure 1. Proposed Model

4.2 Proposed Algorithms Testing

Decision tree classifiers create trees from observational data. This file has branches attached to the branches, which is its main purpose for making leaves. To build an accurate decision tree model, input data (XTrain) and corresponding output data (YTrain) are used for training. Once the model is trained, its performance is evaluated using test data (XTest) and corresponding output data (YTest) by passing them to the score() function, which calculates the model's score. Here, the decision tree model achieved a score of 81.96, indicating its predictive accuracy.

K-Neighbors Classifier is a famous ML algorithm which is simple yet effective non-parametric techinique for classification.By comparing a patient's feature values to those of other patients in a training dataset, KNN can be used to determine whether a patient has heart disease.KNN specifically calculates the distances between the features of the new patient and the features of the training dataset, and based on those distances, chooses the k nearest neighbours.Then, it places the brand-new patient in the class that is most common among their k closest neighbours.

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Random forest distribution is a widely used machine learning method for solving various learning problems such as classification and regression. This algorithm works by creating multiple decision trees, where each tree is trained on a random subset of the data using a technique called "bootstrap aggregation." The resulting trees are then combined to make predictions. In the context of heart attack prediction, the random forest algorithm can be used to build a model that accurately predicts the likelihood of a heart attack based on input data such as age, blood pressure, and cholesterol levels.

$$ni_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)}$$

Support Vector Machine or SVM is a popular machine learning algorithm utilised in the identification of heart disease. Even with highly dimensional data, SVM can effectively handle both linear and non-linear decision boundaries. Compared to other machine learning methods, SVM is also less prone to overfitting.SVM works by finding the optimum hyperplane that separates the data into different classes.Even if the data is not linearly separable in the original feature space, SVM has a kernel trick that can be used to transform the data into a higher dimensional space where it is more separable.

Artificial Neural Networks also known as Multilayer Perceptrons, are an important tool used in machine learning. They are designed to replicate the way humans learn and are considered to be powerful and inspiring models that can help address poor performance issues. As the name implies, ANN systems are inspired by the structure and function of biological neurons and neural networks in the brain.

4.3 Result Analysis

After evaluating the models, we found ANN to be the best performer. So we summarize that that our accuracy is improved due to the increased medical attributes that we have used from the data set we took. It it turns out that we are successfully able to predict, whether the person is diagnosed with Heart disease or not.

Model_IDs	***Algorithms***	***Accuracy* (%)**
1	ANN	90.16
2	KNN	88.52
3	SVM	86.88
4	Logistic Regression	85.24
5	Decision Tree	81.96
6	Naive Bayes	86.88
7	Random Forest	86.88

SECTION - 5 Standards Adopted

Design standards: The system's design follows standard software engineering principles, including modularity, scalability, and maintainability. Similarly, Our design also incorporates best practices for ML model development, such as data pre-processing, feature selection, and model validation.

Testing standards: The system has also undergone rigorous testing at all stages of development, including unit testing, integration testing, and system testing. Testing is automated and test cases are able to cover a wide range of scenarios.

Coding standards: Our code follows industry-standard coding practices, including adherence to a coding style guidelines and proper use of comments and documentation. The code is also modular and organized to facilitate maintenance and future development standards

SECTION - 6
Conclusion and Future Scope

6.1 Conclusion

From this research, after conducting data classification, data pre-processing, data visualization & fitting the model. We finally deployed the model using streamlit via spyder IDE and generated results based on prescribed that whether the person has heart disease or not. By this we finally concluded our project with maximum accuracy score of 0.9016 using the ANN algorithm.

This is a heart disease prediction project based on ML field.While in making this project different data sets are collected.These data sets are again filtered and then suitable machine learning models are selected based on different accuracy.ANN works best (90.16 % accuracy). So we used ANN. At end this model is deployed and gets ready for final use. This research has managed to provide a leap through in predicting heart disease using machine learning & artificial intelligence.

6.2 Future Scope

Current research data provides the best insights into different machine learning-based heart disease diagnostic methods. This work may be modified in the future by adding more features to cardiac data and making it more interactive for users. It can also be done as a mobile application, which reduces computation time and complexity. We will update the system by connecting to the hospital database.

SECTION - 7 References

[1] Loku L., Fetaji B., Krstev A., Fetaji M., Zdravev Z. Using python programming for assessing and solving health management issues South East Eur. J. Sustain. Dev., 4 (1) (2020)

[2] .Overview of machine learning in healthcare Machine Learning Applications using Python, A Press, Berkeley, CA (2019), pp. 1-11

[3] Guleria P., Sood M. Intelligent learning analytics in healthcare sector using machine learning Machine Learning with Health Care Perspective, Springer, Cham (2020), pp. 39-55

[4] Ali, F., Hasan, B., Ahmad, H., Hoodbhoy, Z., Bhuriwala, Z., Hanif, M., et al.(2021). Protocol: Detection of subclinical rheumatic heart disease in children using a deep learning algorithm on digital stethoscope: A study protocol. BMJ Open 11:e044070. doi: 10.1136/bmjopen-2020-044070

[5] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, A. (2017) Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology, 10, 2137-2159.

[6] The IEEE website. [Online]. Available: http://www.ieee.org/

[7] BelItSoft Python in healthcare BelItSoft (2017) Available at-

https://belitsoft.com/custom-application-development-services/healthcare-software-development/python-healthcare , [Accessed on 5th March, 2021]

[8] Panesar A. Machine Learning and AI for Healthcare (1–73) Apress, Coventry, UK (2019)

https://iedu.us/wpcontent/uploads/edd/2020/01/Arjun_Panesar_Machine_Learning_and_AI_for_Health-iedu.us_.pdf , [Accessed on 5th March, 2021]

HEART DISEASE PREDICTION

ANKIT ANSHUMAN MOHAPATRA
2006006

Abstract: This model emphasizes on the development & deployment of an AI-based heart disease diagnosis system using DL & ML algorithms. This study measures the strength & efficiency of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, which is aimed at detecting heart disease based on the maximum accuracy score of the contributing features.

Individual contribution and findings: My role in this project involves data pre-processing, Exploratory Data Analysis, Data Visualization, Fitting Models and deploying it via a web application using Streamlit framework. I had also explored the University of California Data Set, which is a robust data set with strong contributing features that has helped us to generate a optimum accuracy of 90.16% using ANN algorithm and I deployed it via Spyder IDE on a web application. I have also implemented and incorporated several visualization techniques such as heat map, scatter plot, histogram, pie chart & bar graph using seaborn and matplotlib python libraries.

Individual contribution to project report preparation: After exploring multiple research papers, I have analyzed the problem statement, it's visualization and design constraints along with the system architecture. I have also proposed the methodology, procedure and the implementation of the model with a flow chart. Along with it, I have incorporated the model with the UML diagram for heart disease diagnosis process till further deployment.

Individual contribution for project presentation and demonstration: In the PPT portion, I have presented the road-map to organize the framework for the PPT. I have also proposed and incorporated various visualization techniques, that I have used during the implementation, testing & deployment of the heart diagnosis model.

Full Signature of Supervisor: ……

Full signature of the student: ……..

HEART DISEASE DETECTION

RAJDEEP KUNDU

2006514

Abstract: This model emphasizes on the development & deployment of an AI-based heart disease diagnosis system using DL & ML algorithms. This study measures the strength & efficiency of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, which is aimed at detecting heart disease based on the maximum accuracy score of the contributing features.

Individual contribution and findings: I have used KNN and SVM for heart disease prediction based on various features, such as age, sex, blood pressure, cholesterol levels, and other medical history.My results showed 88.52% accuracy in KNN and 86.88% accuracy in SVM.However, the performance of KNN and SVM can be influenced by various factors such as the choice of hyperparameters and the quality of features used in the model.In conclusion, KNN and SVM have shown promising results in heart disease prediction based on various features. However, the performance of these algorithms can be influenced by various factors and further research is required to optimize their performance in this domain.

Individual contribution to project report preparation: I have made the Introduction and the Literature part and also used machine learning algorithms KNN and SVM.

Individual contribution of project presentation and demonstration: I have made slides on KNN and SVM for project presentation.

Full Signature of Supervisor: ……

Full signature of the student: ……..

HEART DISEASE DETECTION

ASHUTOSH YASH
2006518

Abstract:

This model emphasizes on the development & deployment of an AI-based heart disease diagnosis system using DL & ML algorithms. This study measures the strength & efficiency of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, which is aimed at detecting heart disease based on the maximum accuracy score of the contributing features.

My contribution and findings:

Heart disease is a major cause of morbidity and mortality worldwide. To improve risk prediction, i explored the use of machine learning algorithms such as Artificial Neural Networks (ANN) and Naive Bayes Theorem. i trained both algorithms on a dataset of patient information including age, sex, blood pressure, cholesterol levels, and other risk factors.
My results showed that ANN outperformed Naive Bayes and all other algorithms used in predicting the likelihood of heart disease. i optimized the ANN's hyperparameters using cross-validation and achieved an accuracy of 91.16%, compared to 86.88% for Naive Bayes.
My findings suggest that ANN is a powerful tool for improving heart disease risk prediction. Its ability to learn complex relationships between input features and output targets makes it well-suited for handling non-linear relationships between variables and detecting patterns in data. However, it's important to note that the choice of algorithm can depend on the specific dataset and problem at hand. By exploring different machine learning algorithms and comparing their performance, we can improve our understanding of the underlying patterns in data and ultimately, improve patient outcomes.

Full Signature of Supervisor: …….

Full signature of the student: ……..

HEART DISEASE DETECTION

ABHIJEET KUMAR
2006393

Abstract: This model emphasizes on the development & deployment of an AI-based heart disease diagnosis system using DL & ML algorithms. This study measures the strength & efficiency of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, which is aimed at detecting heart disease based on the maximum accuracy score of the contributing features.

Contribution and findings: To implement my part of the work, I first break down the complex algorithms into smaller and more manageable parts. Then I focused on understanding each part and creating a clear and concise explanation for each one. Once I had completed this task, I worked with other members of the group to ensure that our explanations were consistent and easy to understand. During the implementation of my part of the project, I gained valuable technical experience in the field of machine learning. I learned how to apply these algorithms to real-world data sets and how to interpret the results. I also gained a deeper understanding of how to evaluate the performance of these algorithms and how to improve their accuracy. Overall, my contribution to the Heart Disease Detection project was essential in helping our team achieve our goal of developing an accurate and reliable system for detecting heart disease.

Individual contribution to project report preparation: I have contributed in the standards adopted in the report following the Design, Coding and Testing aspects along with that I also proposed and analyzed the results of the models. I also used machine learning algorithms Random Forest and Decision Tree.

Contribution for project presentation and demonstration: My role in preparing presentation and demonstration were instrumental in presenting a clear and concise overview of the Random Forest and Decision Tree algorithms and their applications in heart disease detection.

Full Signature of Supervisor: ……

Full signature of the student: ……..

Heart Disease Pred

ORIGINALITY REPORT

21%	20%	11%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	5%
2	www.researchgate.net Internet Source	3%
3	www.scirp.org Internet Source	2%
4	Submitted to Swinburne University of Technology Student Paper	2%
5	www.worldleadershipacademy.live Internet Source	2%
6	Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, Alamgir Hossain. "An artificial intelligence model for heart disease detection using machine learning algorithms", Healthcare Analytics, 2022 Publication	2%
7	www.frontiersin.org Internet Source	1%