HEART DISEASE PREDICTION

A PROJECT REPORT

Submitted by,

M. Bhuvaneshwar	**-20201COD0024**
Y. Remanth Kumar	**-20201COD0019**
C. Charan Sai	**-20201COD0025**
Y. Gnaneshwar Reddy	**-20201COD0013**

Under the guidance of,

Mr. PAJANY M

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER ENGINEERING [Data Analytics]

At

PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Project report **"HEART DISEASE PREDICTION"** being Submitted by M. Bhuvaneswar 20201COD0024, Y.Remanth Kumar 20201COD0019, C.Charan Sai 20201COD0025, Y. Gnaneshwar Reddy 20201COD0013 in partial fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Engineering [Data Analytics] is a bonafide work carried out under my supervision.

10/1/2024

Mr. PAJANY M
Assistant Professor
School of CSE
Presidency University

11-01-24

Dr.Gopal Krishna shyam
Professor & HoD
School of CSE
Presidency University

Dr. C. KALAIARASAN
Associate Dean
School of CSE&IS
Presidency University

10/1/24

Dr. SHAKKEERA L
Associate Dean
School of CSE&IS
Presidency University

Dr. SAMEERUDDIN KHAN
Dean
School of CSE&IS
Presidency University

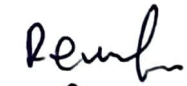
PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **HEART DISEASE PREDICTION** in partial fulfilment for the award of Degree of **Bachelor of Technology** in **Computer Engineering [Data Analytics]**, is a record of our own investigations carried under the guidance of **Mr. PAJANY M, Assistant Professor, School of Computer Science and Engineering, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

Name	Roll. No.	Signature
M. Bhuvaneshwar	20201COD0024	
Y. Remanth Kumar	20201COD0019	
C. Charan Sai	20201COD0025	
Y. Gnaneshwar Reddy	20201COD0013	

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected Dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. Kalaiarasan C and Dr. Shakkeera L,** School of Computer Science & Engineering and Information Science, Presidency University and **Dr. Gopal Krishna Shyam**, Head of the Department, School of Computer Science and Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Mr. Pajany M, Assistant Professor**, School of Computer Science and Engineering, Presidency University for his inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud, Dr. Mrutyunjaya MS** and also the department Project Coordinators **Mrs. Yogeetha B.R, Ms. Sudha P, Dr. Sasidhar Babu,** School of Computer Science and Engineering

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

M. Bhuvaneshwar

Y. Remanth Kumar

C. Charan Sai

Y. Gnaneshwar Reddy

ABSTRACT

An increasing dependence on sophisticated classification and recognition systems is reflected in the expanding use of machine learning in medical diagnostics. These technologies are essential for enabling the early identification of potentially fatal illnesses, which in turn leads to a notable increase in patient survival rates. Given that heart disease is one of the major causes of death worldwide, taking preventative action is imperative. Because younger people are experiencing heart attacks at an increasing rate, a method for early symptom detection must be developed in order to prevent negative consequences.

A practical and trustworthy prediction system is desperately needed, as it is not feasible for the general public to undertake expensive and frequent testing like ECGs. We suggest applying machine learning methods and algorithms, such as XGB Classifier, KNN, SVC, Random Forest, Decision Tree, and Logistic Regression, in response. With an intuitive front-end interface, this system seeks to estimate the risk of heart disease depending on parameters entered by the user. The objective is to provide people with an affordable, easily-accessible technology for proactive health monitoring, which will aid in early intervention and better cardiovascular results.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

People in the modern world struggle with extreme stress and anxiety as a result of their hectic schedules and regular assignments. Furthermore, there is growing concern about people who become addicted to long-term habits like smoking cigars or drinking gutka, which can result in a number of chronic illnesses like cancer, heart disease, liver issues, and kidney failure. For well-known physicians, treating and curing these chronic illnesses is a major challenge.

IT specialists have taken notice of this new difficulty and are assisting in the early detection and treatment of such diseases. Every person is different from the next in terms of appearance, behavior, and blood pressure and pulse rate readings. Medical professionals typically define a healthy blood pressure range of 120/80 to 140/90 mmHg and a healthy pulse rate of 60 to 100 bpm.

The health sector uses a range of machine learning methods and tools in the market today to forecast chronic illnesses. In spite of these efforts, scientists have found certain shortcomings and are looking for more precise predictive algorithms to identify chronic illnesses in people early on and potentially save lives. Thus, based on user-input parameters at the front end, we propose a system that uses machine learning techniques and algorithms, such as XGB Classifier, KNN, SVC, Random Forest, Decision Tree, and Naïve Bayes

CHAPTER-2

LITERATURE SURVEY

Authors	Title	Conference/Journal	Algorithms Used	Key Findings
P. Kola Sujatha and K. Mahalakshmi[1]	Performance evaluation of supervised machine learning...	2020 IEEE International Conference for Innovation in	Naïve Bayes	Random Forest outperforms other algorithms with an accuracy of 83.52%, F1-Score 84.21%, AUC 88.24%, Precision 88.89%.
Pahulpreet Singh Kohli and Shriya Arora[2]	Application of machine learning in disease prediction	2018 4th International Conference on Computing	Decision Tree classification algorithms	Machine learning can be used to detect diseases early on, using three distinct disease databases.
Abderrahmane Ed-Daoudy and Khalil Maalmi[3]	Real-time machine learning for early detection of...	2020 5th Signal Processing and Communications	KNN	Proposes a real-time heart disease prediction system using Apache Spark for large-scale distributed computing.
A. Lakshmanarao A. Srisaila, [4]	Heart disease prediction using feature selection and...	2020 IEEE International Conference for Innovation in	Feature selection, ensemble learning techniques	Utilizes feature selection and ensemble learning techniques for heart disease prediction.
Alperen Erdoğan and Selda Guney[5]	Heart disease prediction by using machine learning...	2020 28th Signal Processing and Communications	Support Vector Machine, Random Forest, Decision Tree	Explores the use of machine learning algorithms for heart disease prediction.
Shaik Farzana and Duggineni Veeraiah[6]	Dynamic heart disease prediction using multi-machine...	2020 5th International Conference on Computing	Multi-machine learning techniques	Proposes a dynamic approach to heart disease prediction using multi-machine learning techniques.

TABLE 2.1 LITERATURE REVIEW

CHAPTER-3

RESEARCH GAPS OF EXISTING METHODS

Despite significant progress, there are still areas in need of study and improvement in the field of machine learning-based heart disease prediction. Understanding these gaps is crucial to developing prediction models that are more reliable and accurate. Some unfulfilled gaps and existing methods in the field of heart disease prediction are as follows:

3.1. Limited Variability within the Sets:

1. Research Gap: Many of the datasets currently used for the prediction of heart disease lack diversity in terms of demographics, lifestyle, and geographic representation.
2. Existing methods: Researchers have used popular datasets, such as the Framingham Heart Study and Cleveland Heart Disease datasets. To increase the generalizability of the model, more diverse datasets from different papulation to enhance models.

3.2. Choosing Features and Their Significance:

1. Research Gap: It's still difficult to determine which features are most important for predicting heart disease. Comprehending the attributes that substantially contribute to precise forecasting is crucial for both clinical applicability and model interpretability.

2. Current Approaches: Recursive Feature Elimination (RFE), feature importance derived from tree-based models, and domain knowledge-driven feature selection are a few of the feature selection strategies that have been used. But more reliable and automated techniques are required.

3. Unbalanced Collections:

4. Research Gap: Biased models can result from imbalanced datasets, where one class (such as the presence of heart disease) is noticeably underrepresented.

5. Current Methods: To address class imbalance, methods such as under sampling, oversampling, and the use of synthetic data (SMOTE) have been used. Nevertheless, more research is required to find the best strategy for imbalanced datasets related to heart disease.

3.3. Unbalanced Collections:

1. Research Gap: Biased models can result from imbalanced datasets, where one class (such as the presence of heart disease) is noticeably underrepresented.
2. Current Approaches: To address class imbalance, methods such as under sampling, oversampling, and the use of synthetic data (SMOTE) have been used. Nevertheless, more research is required to find the best strategy for imbalanced datasets related to heart disease.

3.4. Model Interpretability:

1. Research Gap: Interpretability issues prevent many machine learning models—especially complex ones like ensemble methods—from being widely used in clinical settings.
2. Current Approaches: Interpretability strategies have been investigated, including model-agnostic approaches, LIME, and SHAP values. Nonetheless, creating models that strike a balance between interpretability and complexity is a never-ending task.

3.5. Temporal Elements:

1. Research Gap: The onset of heart disease is a dynamic process that is impacted by time. Current models frequently ignore the data's temporal dimensions.
2. Current Techniques: To capture temporal trends, time-series analysis techniques and longitudinal studies have been investigated. To create models that take into account how heart disease risk factors are changing, more research is necessary.

CHAPTER-4
PROPOSED METHODOLOGY

Creating a methodology that is effective for predicting heart disease requires a methodical approach that takes into account various aspects such as data collection, preprocessing, feature engineering, model selection, evaluation, and validation. A suggested methodology for a project to predict heart disease is provided below:

4.1. Problem Definition and Data Collection:

The goal of the heart disease prediction project is to develop a machine learning model, employing the Random Forest algorithm, that accurately predicts the likelihood of heart disease based on diverse input features such as lifestyle, health history, and demographics. The intended audience includes healthcare professionals, individuals interested in assessing their own risk, and researchers studying cardiovascular diseases. The project aims to achieve high accuracy, identify influential factors, and provide a user-friendly interface, with strict adherence to data privacy regulations and a clear boundary against replacing professional medical advice or diagnosis.

4.2. Data Preprocessing:

Managing missing data involves either removing or imputed values from the dataset.

Data cleaning: Take care of the data's errors, inconsistencies, and outliers.

Normalize and standardize: Scale numerical characteristics to guarantee consistency.

Code variables that are categorical: Utilize methods such as one-hot encoding to translate categorical variables into numerical representations.

4.3. Exploratory Data Analysis (EDA):

To comprehend feature distribution, spot correlations, and learn more about possible relationships between variables, do exploratory data analysis (EDA).

To identify patterns and trends in data, visualize it with graphs and charts.

4.4. Feature Engineering:

Choose pertinent features: Determine which variables have the greatest influence by using features' importance, domain expertise, or feature selection methods.
Make additional features if needed: Extrapolate features that could improve the model's ability to predict the future.

4.5. Training and validation

Divide the dataset into sets for training and validation. Set aside some data for the model's training and some for its validation.
Put cross-validation into practice: Make use of methods such as k-fold cross-validation to evaluate model performance with reliability.

4.6. Hyperparameter Tuning:

To improve model performance, optimize model parameters with methods like grid search or random search.

4.7. Validation and Deployment:

Test the model using a different dataset: Evaluate the model's ability to be generalized.
Use the model in an actual clinical setting if it performs well.
Establish a mechanism for ongoing model monitoring and updating so that it can adjust to modifications in patient demographics or data distribution.

4.8. Documentation:

Record every step of the process, including the data preprocessing stages, the reasoning behind feature selection, the model selections, and the performance metrics.

CHAPTER-5

OBJECTIVES

Addressing critical gaps in machine learning-based heart disease prediction is imperative for improving model reliability and accuracy. Notably, current datasets lack diversity in demographics and lifestyle, emphasizing the need for more varied datasets to enhance model generalizability. Feature selection remains challenging, and while methods like Recursive Feature Elimination are utilized, there is a demand for more reliable automated techniques. Dealing with imbalanced datasets, a common issue, requires further investigation into the most effective strategies beyond current methods. Model interpretability is another gap, with a need for a better balance between complexity and interpretability, especially for complex models like ensembles. Additionally, incorporating the temporal dimension in models is essential, recognizing heart disease as a dynamic process impacted by time. Efforts should focus on developing models that consider the evolution of risk factors over time to capture the dynamic nature of the disease.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

Data Flow Diagram

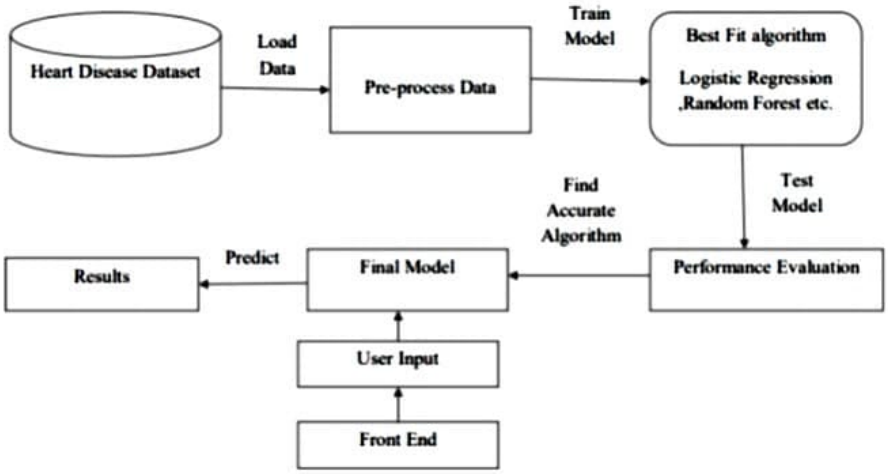


Fig 6.1. Data Flow Diagram

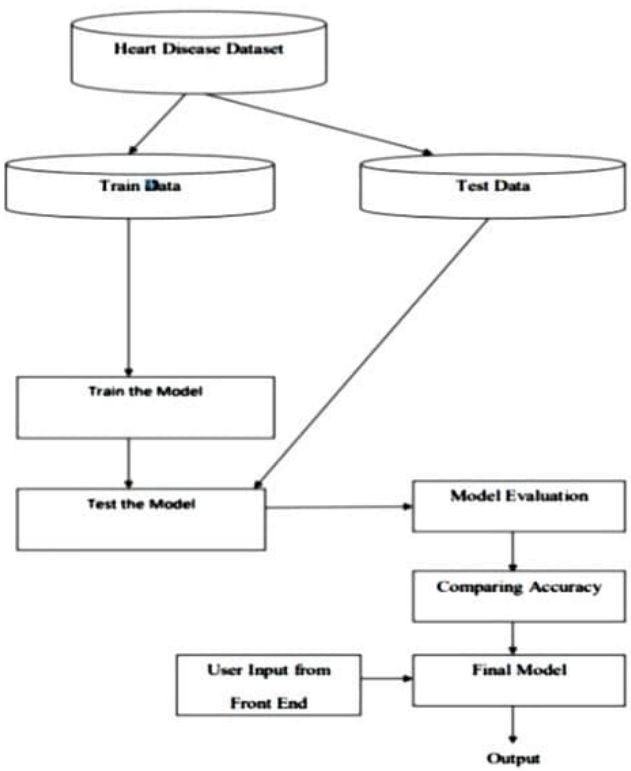


Fig 6.2. System architecture

In the heart disease prediction system, the model evaluation process is a critical step to ensure its reliability and accuracy. Initially, a diverse dataset encompassing demographics, lifestyle, and health history is acquired. This dataset is then split into training and testing subsets to facilitate model training and evaluation. The training phase involves utilizing algorithms like Random Forest to train the model on the training dataset, adjusting parameters for optimal performance. Subsequently, the model is tested on the reserved dataset to assess its ability to generalize to new, unseen data.

During the evaluation, various metrics such as accuracy, precision, recall, and F1 score are calculated to quantify the model's performance. The comparison of these metrics aids in gauging the effectiveness of the model in predicting heart disease. Accuracy, representing the overall correctness of predictions, is a key metric, but other measures provide a more nuanced understanding, especially considering potential class imbalances in the dataset.

The output of the model evaluation phase includes not only quantitative metrics but also qualitative insights into the model's strengths and potential areas for improvement. These insights are crucial for refining the model and enhancing its predictive capabilities. The iterative nature of this process ensures continuous improvement based on feedback from the testing phase.

The user interface (UI) plays a pivotal role in this entire process, serving as the gateway for users to input their data securely. The UI should not only collect user inputs effectively but also present the model's predictions in a clear and understandable format. User feedback from the UI is invaluable for refining the model further, addressing any discrepancies, and making the system more user-friendly. This closed feedback loop between users and the system ensures ongoing optimization and adaptation to user needs.

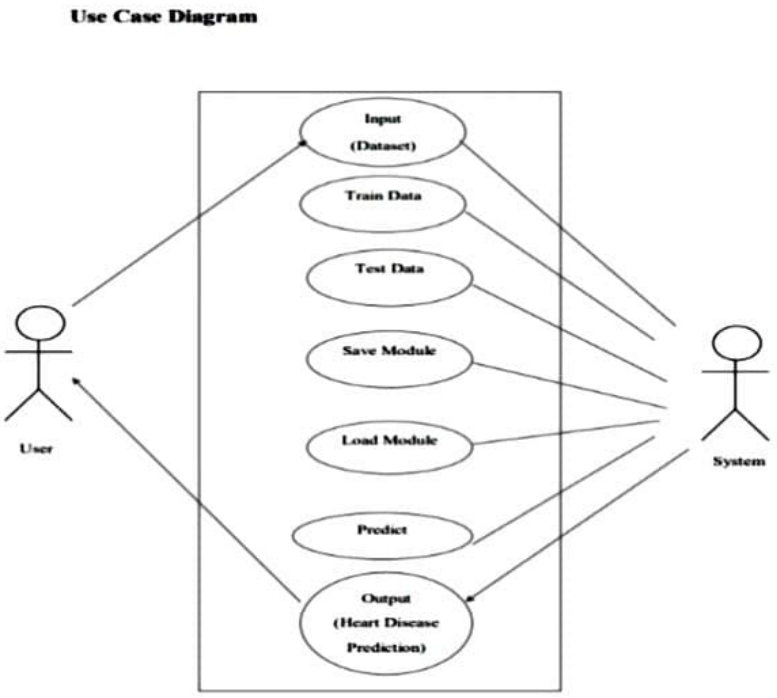


Fig 6.3. Use Case Diagram

The heart disease prediction system begins by collecting relevant data from the user. Once the data is gathered, the system initiates an analysis to extract meaningful insights. Subsequently, the collected data undergoes a preprocessing phase where missing values are addressed, and any necessary adjustments are made to enhance the overall quality of the information. The system then proceeds to build a predictive model using machine learning algorithms, incorporating techniques such as Logistic Regression, Random Forest Classifier, KNN, SVC, Naïve Bayes, Decision Tree Classifier, and XGB Classifier.

Upon constructing the model, a critical step involves evaluating and testing its accuracy. This ensures that the predictive capabilities of the model are reliable and effective. The user actively participates in the decision-making process by finalizing the model with the best accuracy among the evaluated algorithms. This user-driven selection is crucial for tailoring the system to specific preferences or requirements.

Once the model is finalized, the user can provide new data to the system. The predictive model then utilizes the established algorithms to analyze the input and predict the likelihood of heart disease. The results are subsequently displayed to

the user, offering valuable insights into the individual's potential risk. In essence, the system seamlessly integrates user input, sophisticated analysis, and machine learning capabilities to provide a user-friendly and accurate tool for heart disease prediction.

Activity Diagram

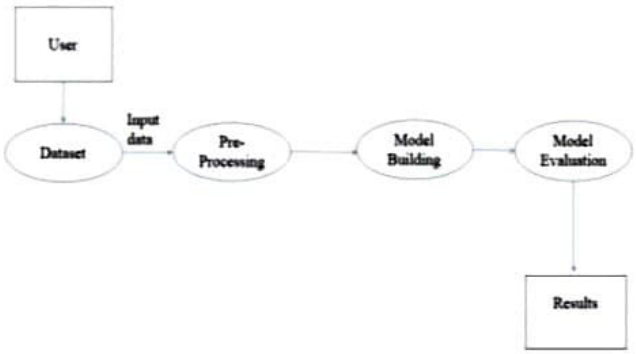


Fig 6.4. Activity Diagram

In the workflow of the heart disease prediction system, users initiate the process by providing a dataset to the system. This dataset is then subjected to a preprocessing stage aimed at enhancing the overall accuracy of the subsequent predictive model. The preprocessing step involves handling missing values, addressing outliers, and making necessary adjustments to ensure the dataset is optimized for model building.

Once the dataset is appropriately preprocessed, the system proceeds to construct the heart disease prediction model using a variety of algorithms. These algorithms, which may include Logistic Regression, Random Forest Classifier, KNN, SVC, Naïve Bayes, Decision Tree Classifier, and XGB Classifier, contribute to the creation of a robust and versatile predictive model.

After model construction, an evaluation phase ensues to assess the performance of each algorithm. The system determines the algorithm that yields the highest accuracy, and this particular model is then finalized for use. The finalization step involves selecting the model with the best accuracy, ensuring that it is well-suited for the specific dataset and user requirements.

Subsequently, the user can leverage the finalized model to predict results based on new or existing data. The predictive capabilities of the model provide valuable insights into the likelihood of heart disease, offering a practical and

efficient tool for risk assessment. In summary, this user-centric approach, involving dataset provision, preprocessing, model building, evaluation, and finalization, culminates in a system that empowers users with accurate and informed predictions related to heart disease.

Sequence Diagram

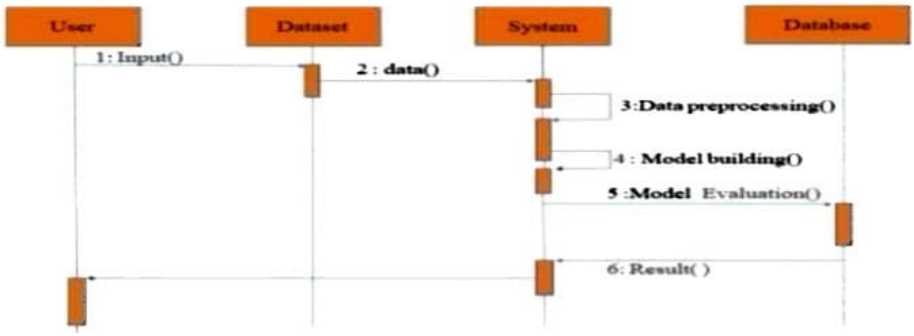


Fig 6.5. Sequence Diagram

The heart disease prediction system operates in a user-friendly manner, beginning with users providing datasets as inputs. The system then efficiently stores the user-provided dataset in its database, ready for subsequent processing. In the preprocessing phase, the stored data undergoes careful handling, addressing missing values and optimizing the dataset for analysis to enhance the accuracy of the subsequent model.

Following preprocessing, the system embarks on the crucial task of model construction. Utilizing various machine learning algorithms, including but not limited to Logistic Regression, Random Forest Classifier, KNN, SVC, Naïve Bayes, Decision Tree Classifier, and XGB Classifier, the system builds a predictive model. This model is subsequently trained using the preprocessed data, enabling it to learn and adapt to patterns within the dataset.

Once the model is constructed and trained, an evaluation process ensues to gauge the performance of each algorithm. The system carefully assesses accuracy metrics, ultimately selecting the algorithm that demonstrates the highest accuracy for finalization. This chosen algorithm becomes the basis for the finalized model, tailored to provide optimal predictive capabilities.

The culminating step involves leveraging the finalized model to predict results based on new or user-provided data. Users can rely on the accuracy of the model

to gain valuable insights into the likelihood of heart disease. In essence, this streamlined process, from dataset input to model finalization, underscores the system's efficacy in providing users with accurate and reliable predictions, contributing to informed decision-making regarding heart disease risk assessment.

Implementation:

In the implementation phase, Python serves as the primary programming language for developing the heart disease prediction system. Python is chosen for its interpreted nature, object-oriented paradigm, and high-level features with dynamic semantics. The language's built-in data structures, combined with dynamic typing and dynamic binding, make it ideal for Rapid Application Development and scripting. Python's simplicity, readability, and extensive standard library contribute to its attractiveness, especially in the context of machine learning (ML) education.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1
61	0	0	145	307	0	0	146	1	1	1	0	3	0

TABLE 6.1 Training Data

Within the Python ecosystem, the Python IDLE (Integrated Development Environment) is utilized as the development environment. IDLE is a dedicated program for software development that integrates various tools tailored for efficient development. These tools include a code editor with features such as

Python IDLE is a Python shell window that provides interactive interpretation, auto-completion, syntax highlighting, smart indentation, and a basic integrated debugger. It is specifically designed to work with Python 3.6.8. Other features include source control capabilities, build, execution, and debugging tools.

The heart disease prediction system's web-based interface is organized using HTML (Hypertext Markup Language). The code that specifies the organization and content of a webpage is called HTML. The language employs elements—such as paragraphs, lists, pictures, and data tables—enclosed in opening and closing tags to organize material. An opening tag, a closing tag, and the content inside make up each HTML element. Extra details about the element are contained in attributes, which are used to provide the content more features. The class attribute in HTML is used to provide non-unique identifiers so that style information can be applied.

An HTML element's content, closing tag, and opening tag define its anatomy. The content of an element is the text or data that is contained within it; the opening and closing tags indicate the start and finish of the element, respectively. Features that aren't present in the content itself, like class, are added to give more details about the element. The attribute name should be separated from the element name by a space, the attribute value should be surrounded in quotation marks, and the attribute name should come after an equal sign.

In summary, the heart disease prediction system implementation combines the power of Python for machine learning algorithms and data processing with HTML for creating a user-friendly web interface, demonstrating the synergy between programming languages and web technologies in developing practical applications.

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT

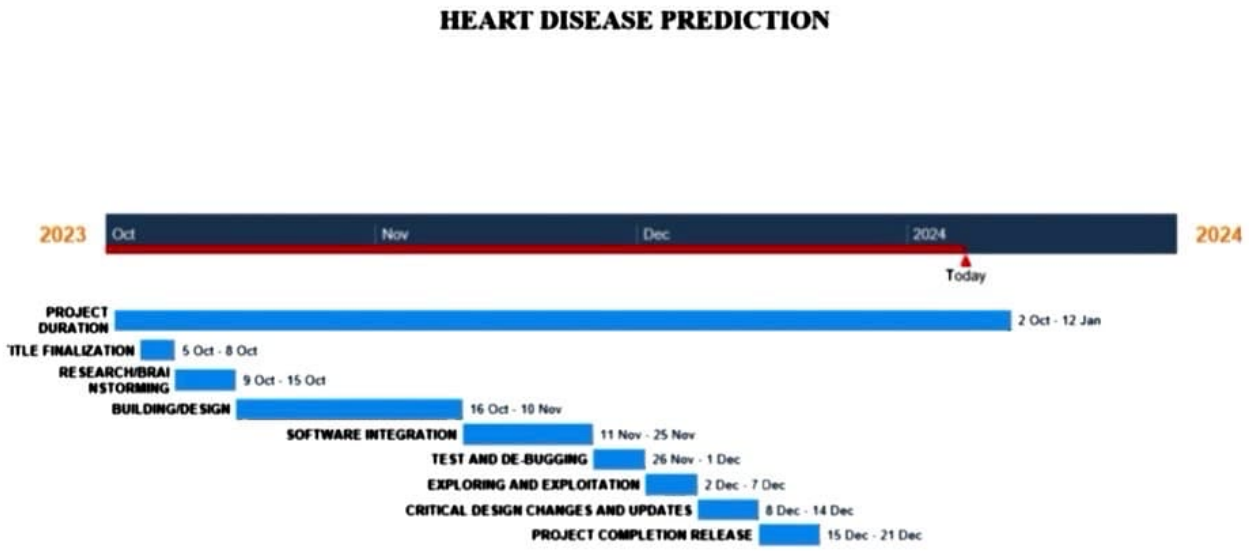


Fig 7.1. Timeline Gantt Chart

Link:
https://online.officetimeline.com/shareable-link?token=YE%2fGVyIrIQ%2fqs96hj25DTmfKEFPmb7i%2fGD4UOpyAzMFJ347llYlONSDsgapqI3qm3n4x3u9O0uW5cSsPIboRAcWjPOGZ38HiA6JxFy7QnI8Y7Pwgqyeh0Bhz6JLb1SN3

CHAPTER-8

OUTCOMES

The heart disease prediction system's adoption has produced a number of noteworthy results. The most important of these is the development of an accurate prediction model, which is accomplished by applying several machine learning algorithms, such as Decision Tree Classifier, XGB Classifier, KNN, SVC, Naïve Bayes, Random Forest Classifier, and Logistic Regression. This guarantees a thorough and solid model that can produce accurate forecasts for cardiac disease.

Additionally, the system boasts a user-friendly interface, thanks to the incorporation of HTML. This enables users to seamlessly interact with the system, providing datasets and visualizing results with ease. The interface enhances the overall user experience, making the system accessible and intuitive.

Moreover, the implementation leverages Python's capabilities for efficient data processing. With its high-level built-in data structures and dynamic semantics, Python proves to be instrumental in handling and processing the datasets effectively. This efficiency contributes to the system's overall performance and reliability.

In conclusion, the outcomes of the heart disease prediction system implementation include the successful development of an accurate model, a user-friendly interface, and efficient data processing capabilities. These achievements collectively contribute to a system that is not only reliable and accurate but also accessible and user-centric in its design and functionality.

CHAPTER-9
RESULTS AND DISCUSSIONS

The endeavor to predict and detect heart disease has long been a formidable challenge for healthcare practitioners. With expensive therapies and operations being the norm for treating heart diseases, the importance of early detection cannot be overstated. The ability to predict heart disease in its early stages holds immense potential for individuals worldwide, allowing them to take proactive measures and prevent the progression of the condition to a more severe state.

We developed and implemented a system utilizing machine learning techniques and algorithms in response to this pressing healthcare issue. Based on a number of parameters that the user enters at the front end, the system is intended to forecast heart disease. Logistic Regression, KNN (K-Nearest Neighbors), SVC (Support Vector Classifier), Random Forest, Decision Tree, XGB Classifier, and Naïve Bayes are the machine learning methods that have been selected. Every algorithm adds distinct skills to the prediction model, guaranteeing a thorough examination of the input parameters..

Upon implementation and testing, our project demonstrated promising results in the prediction of heart disease. The system achieved a commendable accuracy rate of 91.80%, as determined by the Random Forest Classifier. This high level of accuracy is indicative of the effectiveness of the machine learning model in discerning patterns and associations within the dataset, leading to reliable predictions.

In conclusion, the successful implementation of our heart disease prediction system marks a significant step toward addressing the challenges in healthcare related to heart diseases. The achieved accuracy of 91.80%, particularly with the Random Forest Classifier, underscores the potential of machine learning in contributing to early detection and prediction. This system holds promise for empowering individuals to make informed decisions about their health and well-being, paving the way for proactive measures and improved healthcare outcomes. As technology continues to advance, such predictive models have the potential to revolutionize preventive healthcare practices and contribute to a healthier global population.

SLNo	Models	Accuracy
1	Naïve Bayes	90.16%
2	Logistic Regression	85.25%
3	Random Forest	91.80%
4	KNN	65.57%
5	Decision Tree	78.69%
6	SVC	86.89%
7	XGB Classifier	86.89%

TABLE 9.1. Comparison of other algorithms with Random forest

CHAPTER-10

CONCLUSION

In conclusion, our heart disease prediction system harnesses the power of various machine learning algorithms, achieving an impressive 91.80% accuracy, with Random Forest standing out as a key contributor. The multi-algorithmic approach provides a comprehensive analysis of user-entered parameters. The user-friendly front end, developed with Flask, HTML, and pymysql, facilitates easy interaction, empowering users to proactively engage with their health data. This system holds immense potential for revolutionizing early diagnosis and intervention in heart disease cases, exemplifying the symbiosis of healthcare and machine learning. Ongoing refinement and exploration of additional features promise to further enhance the system's predictive capabilities, shaping the future of preventive healthcare practices globally.

References

[1] P. Kola Sujatha and K. Mahalakshmi. Performance evaluation of supervised machine learning algorithms in prediction of heart disease. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–7, 2020.

[2] Pahulpreet Singh Kohli and Shriya Arora. Application of machine learning in disease prediction. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4, 2018.

[3] Abderrahmane Ed-Daoudy and Khalil Maalmi. Real-time machine learning for early detection of heart disease using big data approach. pages 1–5, 04 2019.

[4] A. Lakshmanarao, A. Srisaila, and Srinivasa Tummala. Heart disease prediction using feature selection and ensemble learning techniques. pages 994–998, 02 2021.

[5] Alperen Erdoğan and Selda Guney. Heart disease prediction by using machine learning algorithms. *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2020.

[6] Shaik Farzana and Duggineni Veeraiah. Dynamic heart disease prediction using multi-machine learning techniques. *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–5, 2020.

APPENDIX-A

PSUEDOCODE

BACK END CODE:

```
# for numerical computing
import numpy as np


# for dataframes
import pandas as pd


#for plotting
import matplotlib.pyplot as plt
import seaborn as sns



# Ignore Warnings
import warnings
warnings.filterwarnings("ignore")



# to split train and test set
from sklearn.model_selection import train_test_split



# Machine Learning Models
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from xgboost import XGBClassifier


from sklearn.metrics import accuracy_score




# to save the final model on disk

data=pd.read_csv('heart.csv')
print(data.shape)

info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non-
anginal pain, 4: asymptomatic","resting blood pressure"," serum cholestoral in mg/dl","fasting blood
sugar > 120 mg/dl","resting electrocardiographic results (values 0,1,2)"," maximum heart rate
```

```
achieved","exercise induced angina","oldpeak = ST depression induced by exercise relative to
rest","the slope of the peak exercise ST segment","number of major vessels (0-3) colored by
flourosopy","thal: 3 = normal; 6 = fixed defect; 7 = reversable defect"]


for i in range(len(info)):
   print(data.columns[i]+":\t\t\t"+info[i])


print(data.columns)
print(data.head())
print(data.describe())
print(data.corr())


data = data.drop_duplicates()
print( data.shape )


print(data.isnull().sum())
data=data.dropna()
print(data.isnull().sum())



data["target"].value_counts().plot(kind="bar", color=["salmon","lightblue"])
plt.xlabel("0 = No Disease, 1 = Disease")
plt.title("Heart Disease")
plt.show()


# Create a plot of crosstab
pd.crosstab(data.target, data.sex).plot(kind="bar",
                                        figsize=(10,6),
                                        color=["salmon","lightblue"])
plt.title("Heart Disease Frequency for Sex")
plt.xlabel("0 = No Disease, 1 = Disease")
plt.legend(["Female","Male"])
plt.show()



y = data.target

# Create separate object for input features
X = data.drop('target', axis=1)
```

```
# Split X and y into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y,  test_size=0.2, random_state=0)


# Print number of observations in X_train, X_test, y_train, and y_test
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)


model1= LogisticRegression()
model2=RandomForestClassifier(random_state=285) #285,1673
model3= KNeighborsClassifier(n_neighbors=9)
model4=DecisionTreeClassifier()
model5= GaussianNB()
model6=SVC(kernel='linear',C=10 ,gamma=0.0009)
model7=XGBClassifier()


model1.fit(X_train, y_train)
model2.fit(X_train, y_train)
model3.fit(X_train, y_train)
model4.fit(X_train, y_train)
model5.fit(X_train, y_train)
model6.fit(X_train, y_train)
model7.fit(X_train, y_train)


## Predict Test set results
y_pred1 = model1.predict(X_test)
y_pred2 = model2.predict(X_test)
y_pred3 = model3.predict(X_test)
y_pred4 = model4.predict(X_test)
y_pred5 = model5.predict(X_test)
y_pred6 = model6.predict(X_test)
y_pred7 = model7.predict(X_test)


acc1 = accuracy_score(y_test, y_pred1) ## get the accuracy on testing data
print("Accuracy of Logistic Regression is {:.2f}%".format(acc1*100))


acc2 = accuracy_score(y_test, y_pred2) ## get the accuracy on testing data
print("Accuracy of RandomForestClassifier is {:.2f}%".format(acc2*100))


acc3 = accuracy_score(y_test, y_pred3)  ## get the accuracy on testing data
```

```
print("Accuracy of KNeighborsClassifier is {:.2f}%".format(acc3*100))


acc4 = accuracy_score(y_test, y_pred4) ## get the accuracy on testing data
print("Accuracy of Decision Tree is {:.2f}%".format(acc4*100))


acc5 = accuracy_score(y_test, y_pred5) ## get the accuracy on testing data
print("Accuracy of GaussianNB is {:.2f}%".format(acc5*100))


acc6 = accuracy_score(y_test, y_pred6) ## get the accuracy on testing data
print("Accuracy of SVC is {:.2f}%".format(acc6*100))


acc7 = accuracy_score(y_test, y_pred7) ## get the accuracy on testing data
print("Accuracy of XGB Classifier is {:.2f}%".format(acc6*100))



#from sklearn.externals import joblib
import joblib

# Save the model as a pickle in a file

joblib.dump(model2, 'heart_disease.pkl')

# Load the model from the file
final_model = joblib.load('heart_disease.pkl')


pred=final_model.predict(X_test)


acc = accuracy_score(y_test,pred)# get the accuracy on testing data
print("Final Model Accuracy is {:.2f}%".format(acc*10

scores = [acc1,acc2,acc3,acc4,acc5,acc6,acc7]
algorithms = ["Logistic Regression","Random Forest","KNN","Decision Tree","Naive
Bayes","SVC","XGB Classifier"]


sns.set(rc={'figure.figsize':(15,8)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")

sns.barplot(algorithms,scores)

plt.show()
```

APPENDIX-B

SCREENSHOT

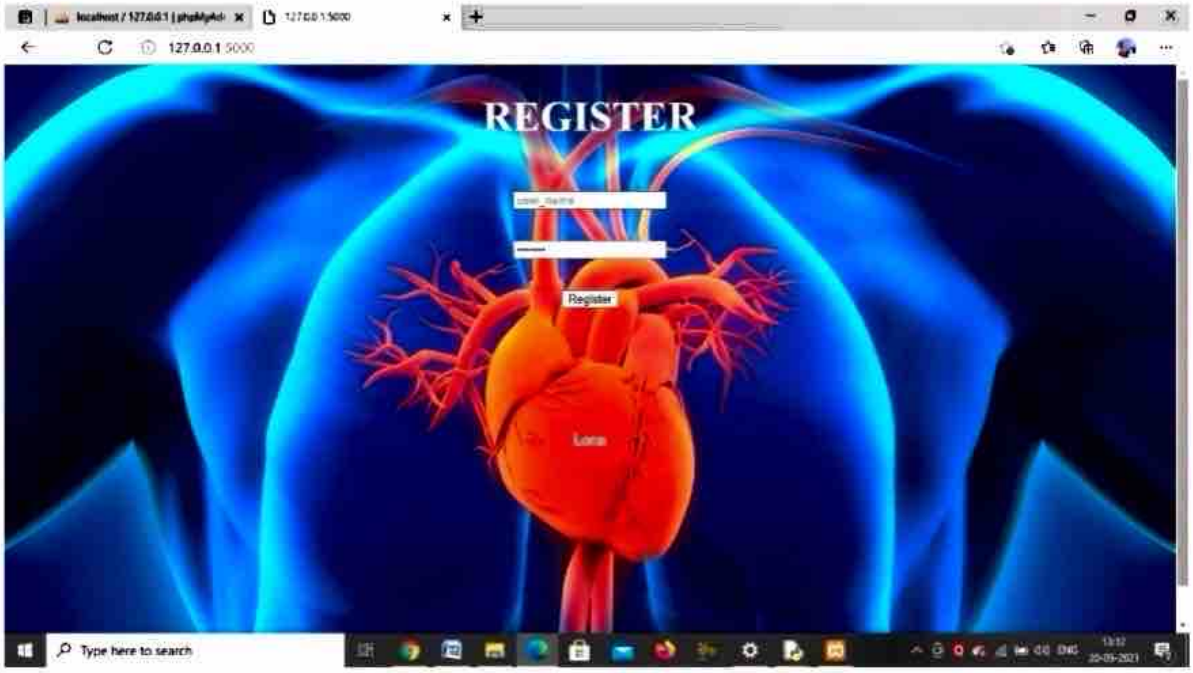


Fig B.1. Register interface

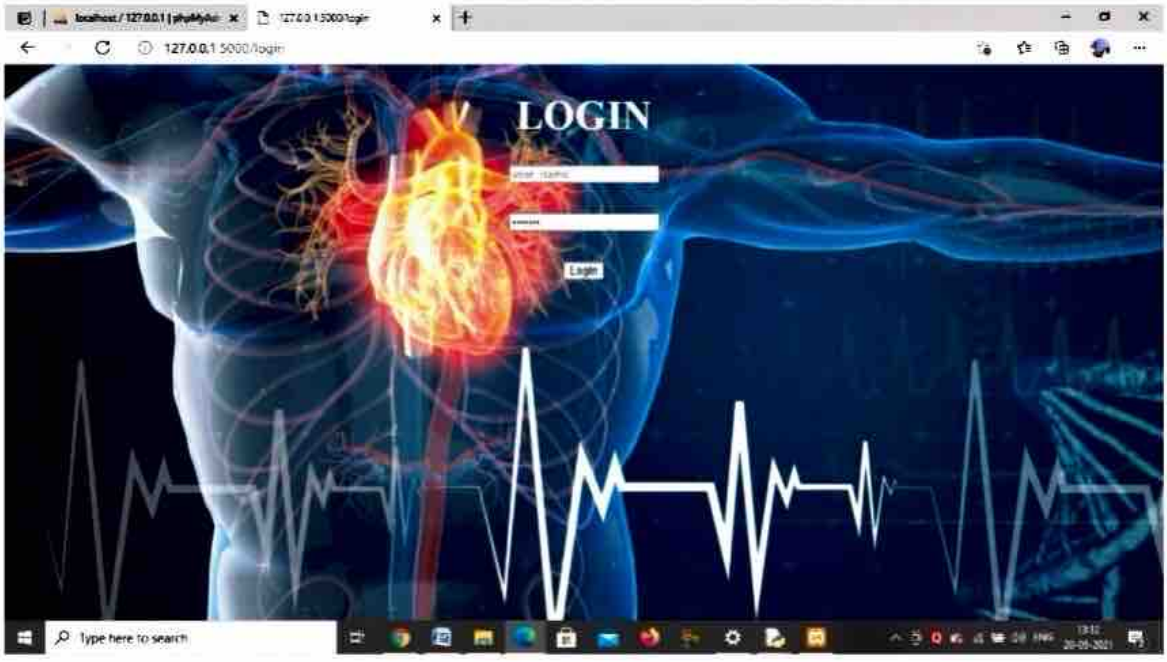


Fig B.2. Login interface

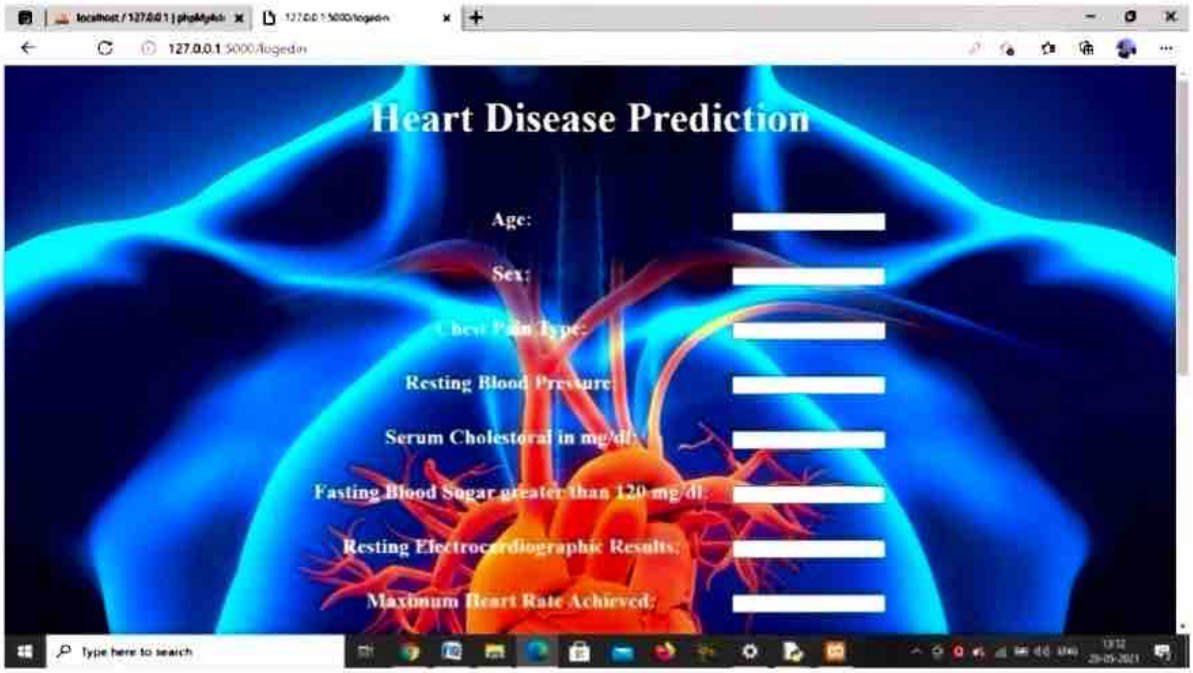


Fig B.3. Prediction interface

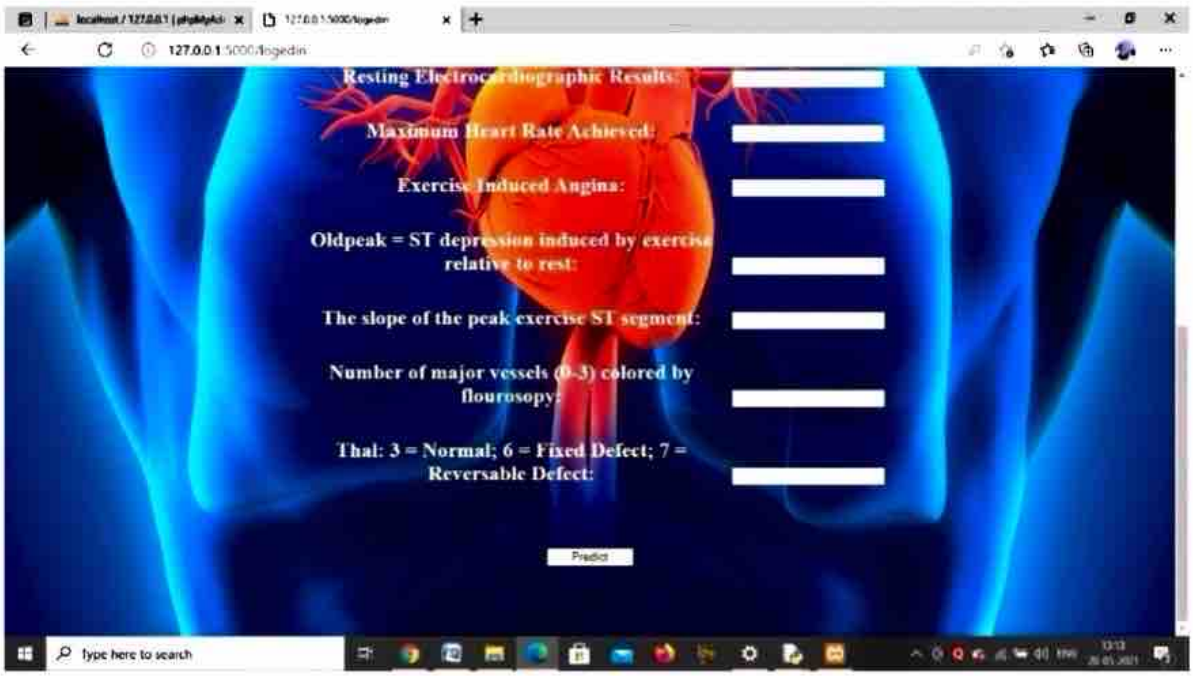


Fig B.4. Prediction interface

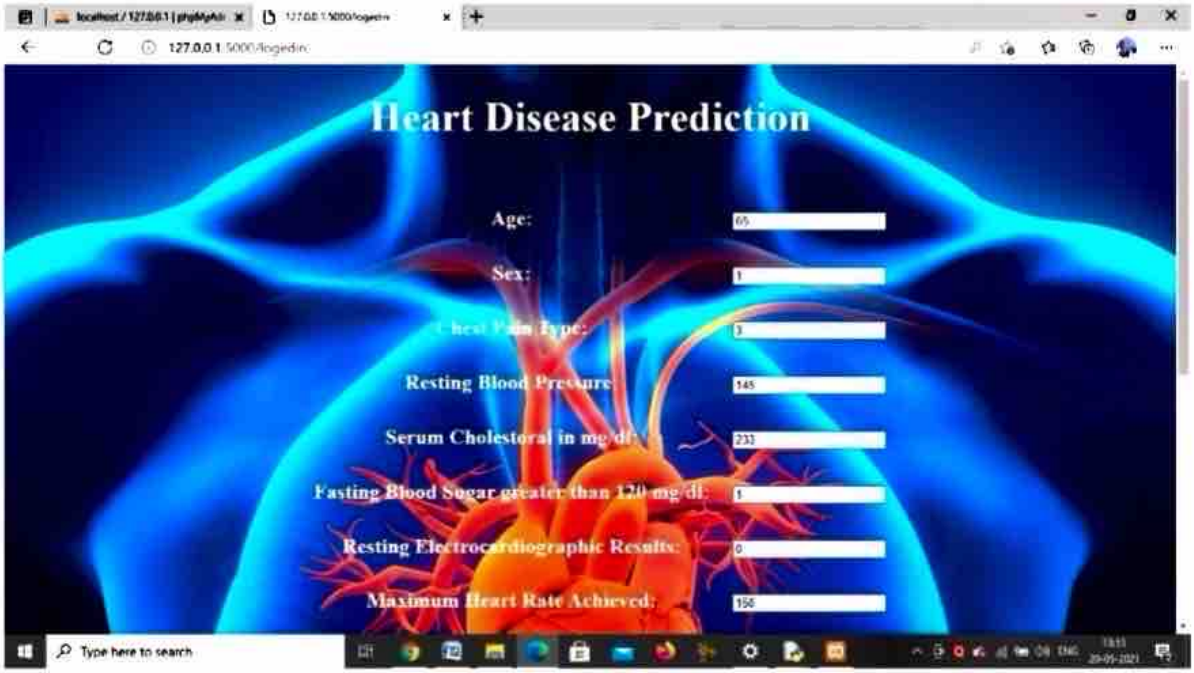


Fig B.5. FINAL OUTPUT OF THE PROJECT

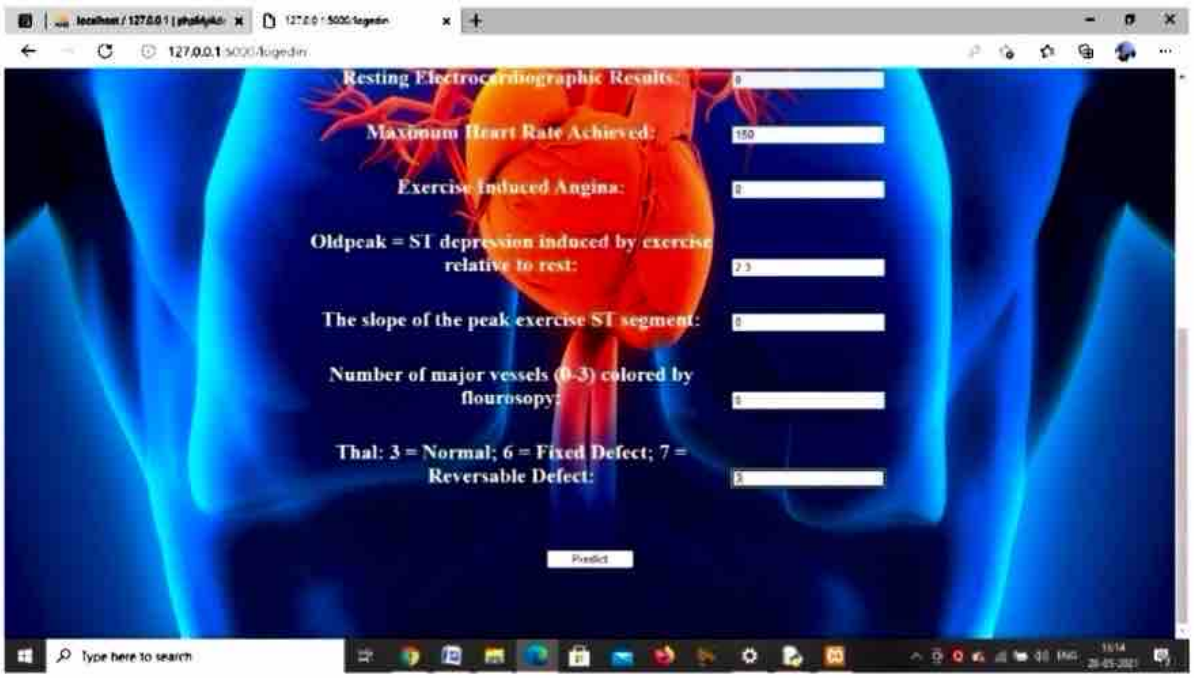


Fig B.6. FINAL OUTPUT OF THE PROJECT

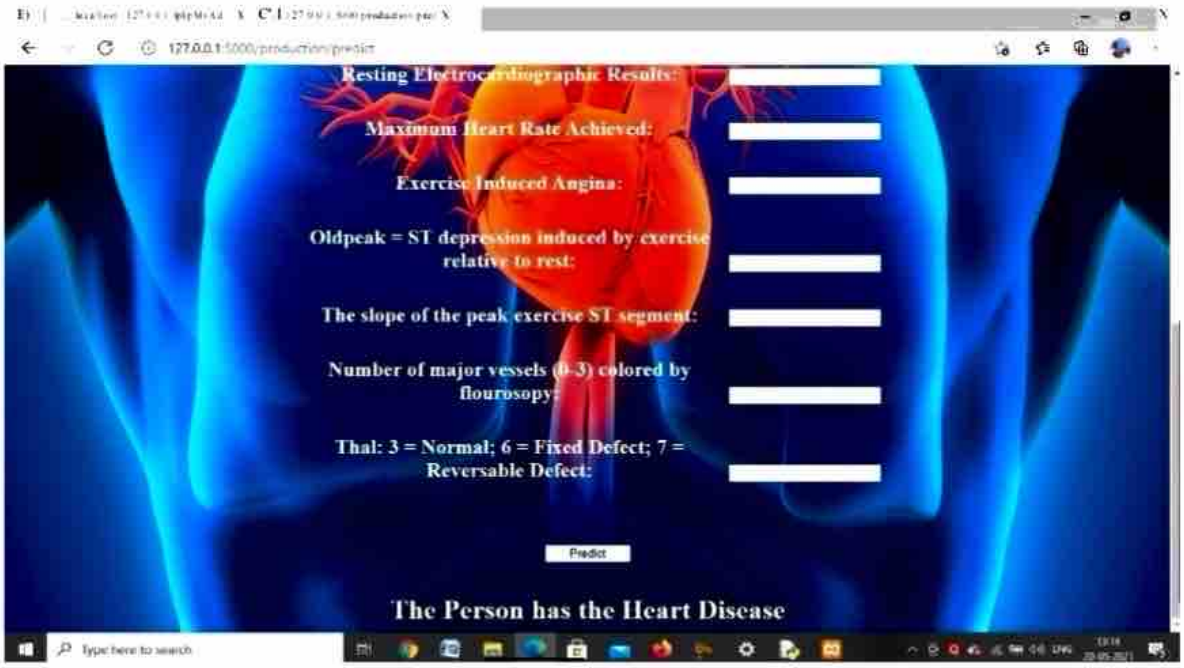
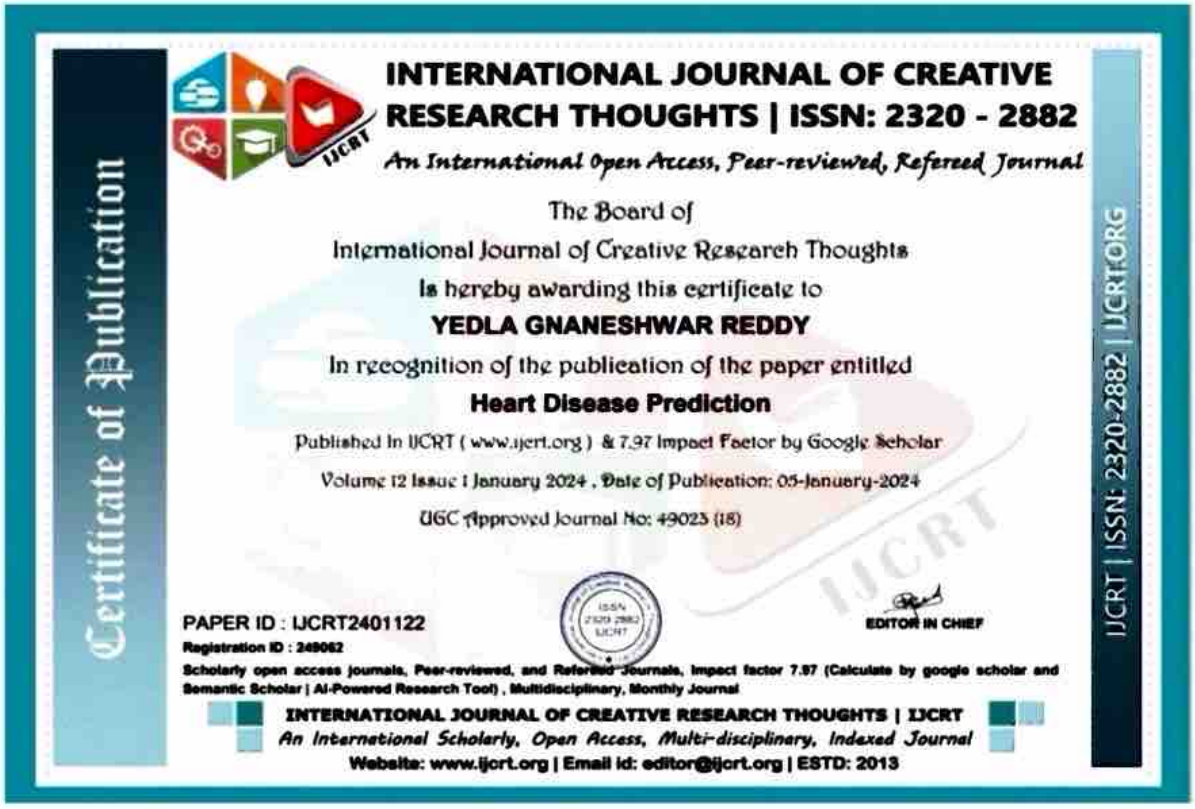


Fig B.7. FINAL OUTPUT OF THE PROJECT

APPENDIX-C

ENCLOSURES



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

YEDLA GNANESHWAR REDDY

In recognition of the publication of the paper entitled

Heart Disease Prediction

Published in IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 12 Issue 1 January 2024 , Date of Publication: 05-January-2024

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2401122

Registration ID : 249062

EDITOR IN CHIEF

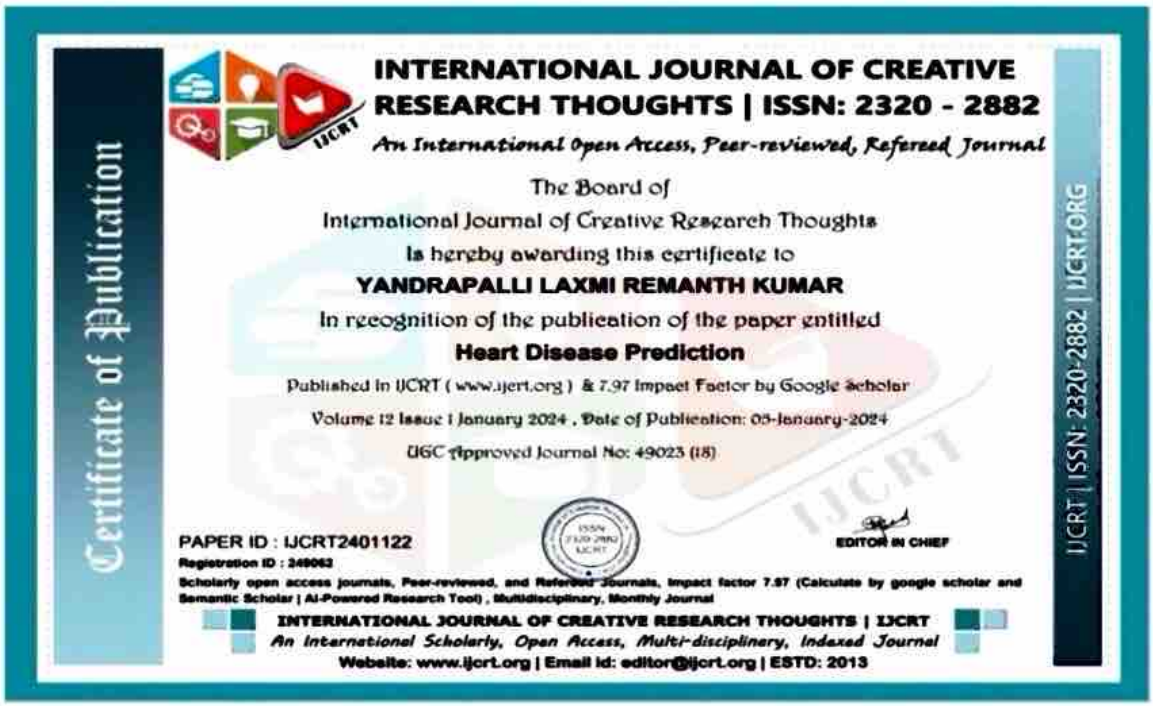
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

YANDRAPALLI LAXMI REMANTH KUMAR

In recognition of the publication of the paper entitled

Heart Disease Prediction

Published in IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 12 Issue 1 January 2024 , Date of Publication: 05-January-2024

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2401122

Registration ID : 249062

EDITOR IN CHIEF

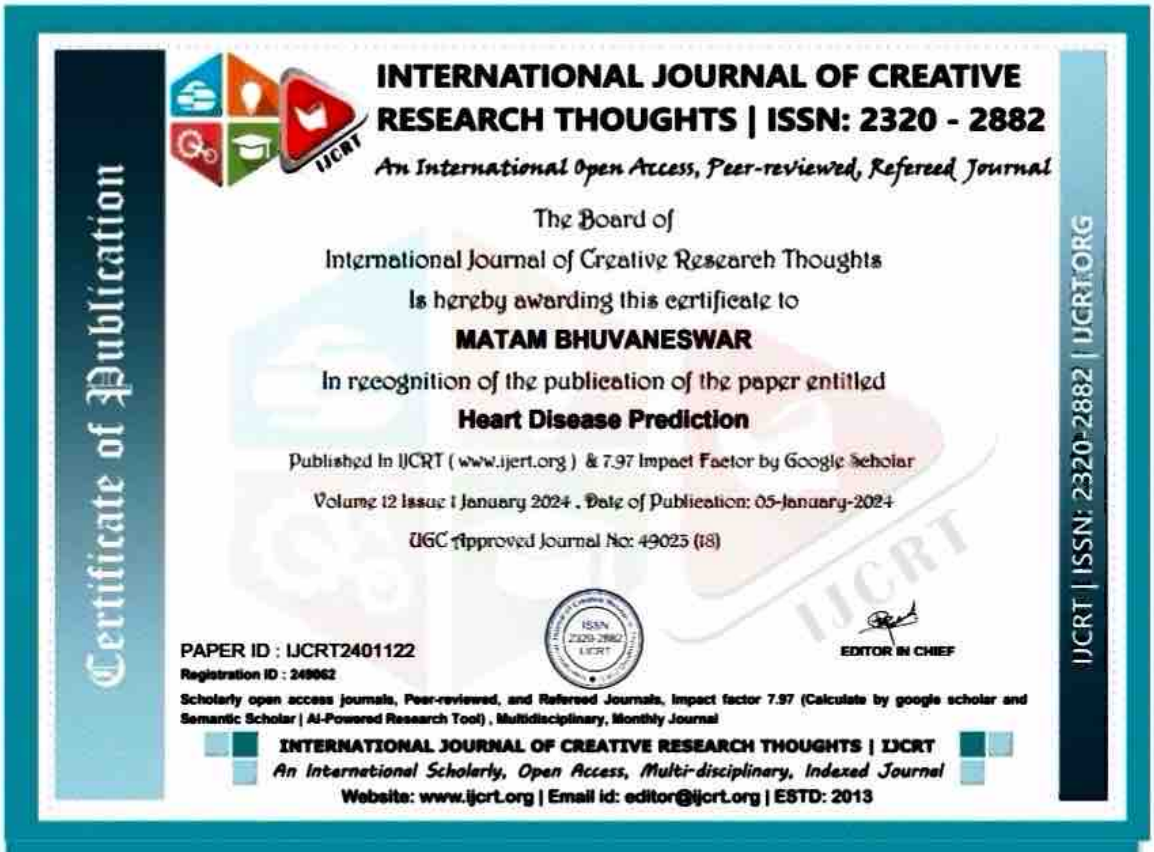
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

MATAM BHUVANESWAR

In recognition of the publication of the paper entitled

Heart Disease Prediction

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 12 Issue 1 January 2024 , Date of Publication: 05-January-2024

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2401122

Registration ID : 249062

EDITOR IN CHIEF

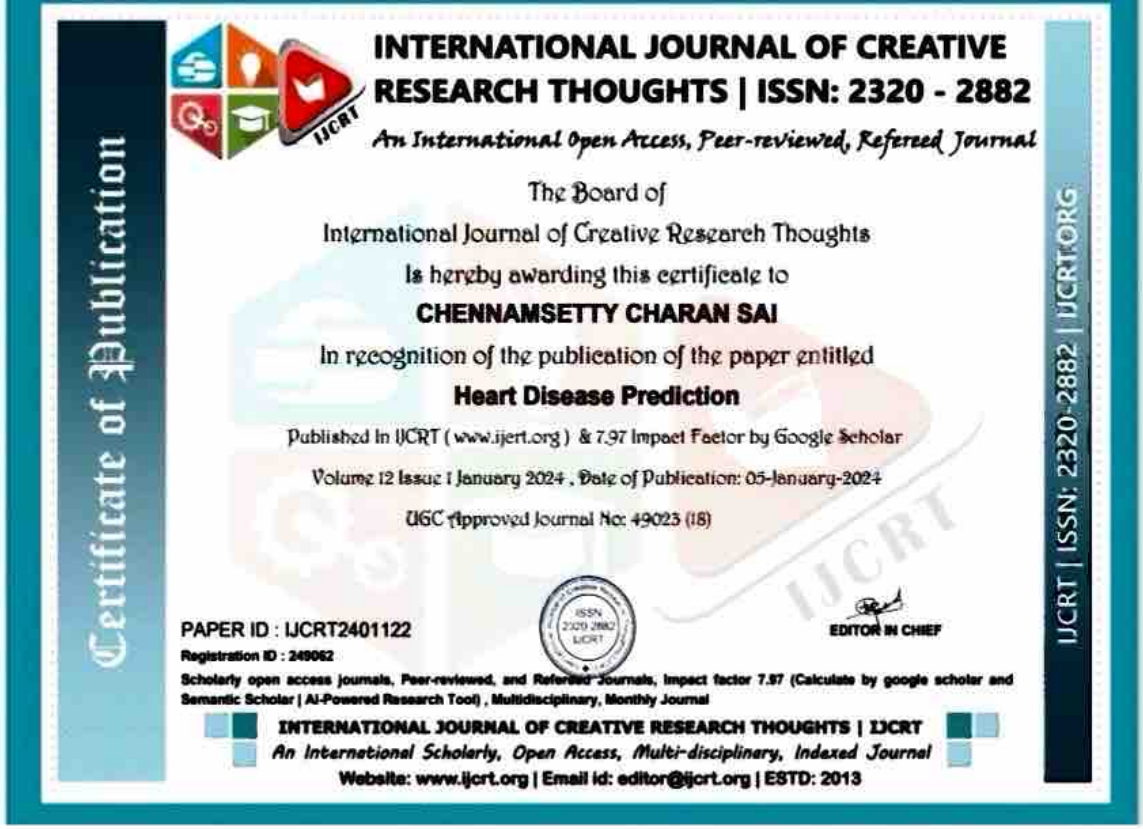
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

CHENNAMSETTY CHARAN SAI

In recognition of the publication of the paper entitled

Heart Disease Prediction

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 12 Issue 1 January 2024 , Date of Publication: 05-January-2024

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2401122

Registration ID : 249062

EDITOR IN CHIEF

Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG

Heart

ORIGINALITY REPORT

21%	15%	13%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of North Texas Student Paper	5%
2	github.com Internet Source	2%
3	Submitted to M S Ramaiah University of Applied Sciences Student Paper	1%
4	www.ijraset.com Internet Source	1%
5	journal.ugm.ac.id Internet Source	1%
6	www.irjmets.com Internet Source	1%
7	hashnode.com Internet Source	1%
8	www.analyticsvidhya.com Internet Source	1%
9	Submitted to University of Essex Student Paper	<1%
10	"ICT for Intelligent Systems", Springer Science and Business Media LLC, 2023 Publication	<1%
11	www.dspace.dtu.ac.in:8080 Internet Source	<1%
12	Adedayo Ogunpola, Faisal Saeed, Shadi Basurra, Abdullah M. Albarrak, Sultan Noman	<1%

The Project work carried out here is mapped to SDG-3 Good Health and Well-Being.

The project work carried here contributes to the well-being of the human society. This can be used for Analyzing and detecting heart disease in the early stages so that the required medication can be started early to avoid further consequences which might result in mortality.