**Multiple Disease Prediction System**

**Using Machine-Learning**

**A Project Report**

Submitted in Partial Fulfillment of the Requirements for

the Award of the Degree of

**Bachelor of Technology**

**In**

**(Electronics and Communication Engineering)**

*Submitted by*

**1.Mohit Kumar 2. Anurag Singh**

2004139 2004102

**3.Ranjeet 4. Aditya Dhar Dwivedi**

2004135 2004132

*Under the supervision of*

**Prof. Bharat Gupta**



**Department of Electronics and Communication Engineering**

**NATIONAL INSTITUTE OF TECHNOLOGY PATNA**

**Patna-800005**

**MAY 2024 (2020-24)**

**Certificate from Supervisor(s)**

This is to certify that **Anurag Singh** (Roll No. 2004102), **Ranjeet** (2004135), **Aditya Dhar Dwivedi** (2004132) and **Mohit Kumar** (2004139) of National Institute of Technology Patna have successfully completed the major project work titled **“Multiple Disease Prediction Using Machine Learning”** in 8th semester of Bachelor of Technology course from **January to May 2024**. This project report is the bonafide work done by his/her/them for the partial fulfillment of the requirements for the award of the degree of B.Tech. in Electronics and Communication Engineering.

**……………………...........**

**Prof. Bharat Gupta**

**(Project Supervisor)**

**DECLARATION AND COPYRIGHT TRASFER**

We **Anurag, Ranjeet, Aditya, and Mohit** Roll No. **2004102, 2004135, 2004132** and **2004139** Enrolment No. **202876, 203052, 203033** and **203066** a registered candidate for Undergraduate Programme (B.Tech/B.Arch.) under department of **Electronics and Communication Engineering** of **National Institute of Technology Patna**, declare that this is my own original work and does not contain material for which the copyright belongs to a third party and that it has not been presented and will not be presented to any other University/ Institute for a similar or any other Degree award.

I further confirm that for all third-party copyright material in my thesis/ dissertation (including any electronic attachments) is “blanked out” third party material from the copies of the thesis/dissertation/book/articles etc; fully referenced the deleted materials and where possible, provided links (url) to electronic sources of the material. I hereby transfer exclusive copyright for this thesis to NIT Patna. The following rights are reserved by the author: a) The right to use, free of charge, all or part of this article in future work of their own, such as books and lectures, giving reference to the original place of publication and copyright holding. b) The right to reproduce the article or thesis for their own purpose provided the copies are not offered for sale.

Signature of the candidate: …………………………………….

Date: ………………………….

**COPYRIGHT © [2024] by NIT Patna. All rights reserved.**

No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission, except in the case of brief quotations embodied in critical scholarly reviews and certain other non-commercial uses with an acknowledgment/reference permitted by the copyright law.

**ACKNOWLEDGEMENT**

We express our sincere gratitude to the Department of Electronics & Communication Engineering for extending all kinds of help while carrying out this work.

We are deeply thankful to our guide, Prof. Bharat Gupta, who spared his valuable time to review our work and gave us constant encouragement to continue our work.

We are grateful to our family, friends and all the well-wishers for their help and support.

**Date: Name:**

|  |
| --- |
| ANURAG SINGH |
| RANJEET |
| ADITYA DHAR DWIVEDI |
| MOHIT KUMAR |

**ABSTRACT**

In recent years the chronic disease like heart disease and diabetes have increased rapidly around the world. In India almost 80 million people are affected by diabetes and the annual number of deaths from heart disease in India is projected to rise from 2.26 million (1990) to 4.77 million (2020). Chronic disease like heart disease and diabetes poses a significant burden on global healthcare systems. Early diagnosis and intervention are crucial for improved patient outcomes and reduced healthcare costs. This project explores the development of a multi-disease prediction system leveraging machine learning algorithms. The project contains mainly five chapters named as 1) Introduction 2) Literature Review 3) Proposed Model 4) Result and 5) Conclusion respectively.

The system focuses on predicting two prevalent diseases: heart disease and diabetes at early stage. We utilize publicly available datasets from Kaggle, a renowned platform for data science. The datasets contain various attributes such as age, sex, glucose level, insulin, BMI etc which are crucial for our machine learning model training.

The heart disease prediction model employs Logistic Regression, a well-established algorithm for classification tasks. Logistic Regression excels at modelling the relationship between features, such as age, blood pressure, and cholesterol, and the binary outcome (presence or absence of heart disease).

We have used various python libraries such as NumPy, pandas, scikit-learn and streamlit. We have used Anaconda and Spyder for our coding purpose.

For diabetes prediction, the system implements a Support Vector Machine (SVM) algorithm. SVMs are powerful tools for identifying complex patterns in data, making them suitable for analysing medical datasets. The SVM model will be trained on features like blood sugar levels, body mass index (BMI), Insulin, age etc to predict the likelihood of diabetes development.

The project encompasses data pre-processing, which involves cleaning, handling missing values, and potentially feature engineering. This ensures the data is suitable for model training and improves prediction accuracy.

This multi-disease prediction system presents a promising approach for early disease detection. The system can potentially serve as a preliminary screening tool, prompting individuals to seek professional medical evaluation. By leveraging machine learning, this project aims to contribute to preventative healthcare strategies and empower individuals to take a more proactive role in managing their health.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| *Certificate* | *ii* |
| *Declaration and Copyright Transfer* | *iii* |
| *Acknowledgement* | *iv* |
| *Abstract* | *v-vi* |
| *Table of Contents* | *vii* |
| **Chapter 1: Introduction** | 1-4 |
| 1.1 Motivation | 1 |
| 1.2 Disease Background | 1-2 |
| 1.3 Relation between disease | 2-3 |
| 1.4 Overview of the Project | 4 |
| **Chapter 2: Literature review** | 5-6 |
| **Chapter 3: Proposed Model** | 7-17 |
| 3.1 Dataset Description | 7-8 |
| 3.2 Algorithms | 8-12 |
| 3.3 Diabetes Prediction Model | 12-14 |
| 3.4 Heart Disease Prediction Model | 14-16 |
| 3.5 Streamlit Integration | 16-17 |
| **Chapter 4: Results** | 18-20 |
| **Chapter 5: Conclusion** | 21 |
| **Chapter 6: Future Scope** | 22-23 |
| **References** |  |

**Chapter 1**

**Introduction**

* 1. **Motivation**

Chronic diseases like diabetes and heart disease are leading causes of morbidity and mortality worldwide. **According to WHO in 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years and an estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.** Therefore, early detection and intervention are crucial for improved patient outcomes and reduced healthcare costs. Traditional diagnostic methods often rely on clinical examinations and laboratory tests, which can be time-consuming and expensive. Machine learning (ML) offers a promising approach for developing non-invasive, cost-effective, and fast disease prediction systems. The main motivation of our project is to develop a machine learning model which can detect these chronic diseases in early stages to reduce the cost and fatality. This project delves into the potential of machine learning (ML) as a proactive tool for multiple-disease prediction system presented in a user-friendly format using Streamlit.

**1.2 Disease Background**

* **Diabetes:** Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels (hyperglycaemia) resulting from defects in insulin secretion, action, or both. It affects over 422 million people globally, with a significant number remaining undiagnosed. Common symptoms include excessive thirst, frequent urination, unexplained weight loss, blurred vision, and slow-healing wounds.
* **Types of diabetes**

There are several types of diabetes, each with its own causes, symptoms, and management strategies. Here are the main types:

**1. Type 1 Diabetes:** This type of diabetes occurs when the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. As a result, the body produces little to no insulin, leading to high blood sugar levels. Type 1 diabetes is typically diagnosed in children and young adults, although it can develop at any age. Management of type 1 diabetes involves lifelong insulin therapy through injections or an insulin pump, along with regular monitoring of blood sugar levels and dietary adjustments.

**2. Type 2 Diabetes:** Type 2 diabetes is the most common form of diabetes, accounting for the majority of cases worldwide. It develops when the body becomes resistant to insulin or when the pancreas fails to produce enough insulin to meet the body's needs. Type 2 diabetes is often associated with lifestyle factors such as obesity, physical inactivity, and poor dietary habits. While it can occur at any age, it is more prevalent in adults, particularly those over the age of 45. Management of type 2 diabetes involves lifestyle modifications, such as diet and exercise, as well as oral medications or insulin therapy in some cases.

**3. Gestational Diabetes:** Gestational diabetes occurs during pregnancy when the body is unable to produce enough insulin to meet the increased demands of pregnancy. It usually develops around the 24th to 28th week of pregnancy and typically resolves after childbirth. However, women who develop gestational diabetes are at an increased risk of developing type 2 diabetes later in life. Management involves blood sugar monitoring, dietary changes, and sometimes insulin therapy to ensure a healthy pregnancy and prevent complications for both the mother and baby.

**4. Other Types of Diabetes:** There are also other, less common types of diabetes, including:

* **Monogenic Diabetes**: Caused by mutations in a single gene and typically diagnosed in childhood or young adulthood.
* **Secondary Diabetes**: Resulting from other medical conditions or medications that affect insulin production or action, such as pancreatic disease or certain medications.
* **Heart Disease:** Heart disease, encompassing various conditions affecting the heart and blood vessels, is the leading cause of death globally. Risk factors include high blood pressure, cholesterol, unhealthy diet, physical inactivity, smoking, and diabetes. Symptoms can vary depending on the specific condition but may include chest pain, shortness of breath, fatigue, and leg pain.

**1.3 Relationship between Diabetes and Heart Disease**

Diabetes is a significant risk factor for heart disease. Chronically high blood sugar levels damage blood vessels and nerves, increasing the risk of atherosclerosis (plaque buildup in arteries) and peripheral artery disease (PAD). Diabetes also contributes to high blood pressure and abnormal blood clotting, further promoting heart disease development.

Death due to cardiovascular disease is 70 percent higher in adults with diabetes than in those without it, the CDC reports. About 32 percent of people with type 2 diabetes have heart disease, according to a 2017 study. At least 68 percent of people with diabetes ages 65 and older will die from some form of heart disease, according to the American Heart Association.

**1.4 Project Overview**

This project focuses on building a system capable of predicting two prevalent conditions - heart disease and diabetes. We have utilized the power of Kaggle, a renowned platform for data science, to access publicly available datasets. Rich in relevant features like demographics, vitals, and lab results, these datasets provide the foundation for training and evaluating our ML models.

The system employs Logistic Regression, a well-established algorithm, for heart disease prediction. It excels at modelling the relationship between features and a binary outcome (presence/absence of heart disease). For diabetes prediction, we utilize Support Vector Machines (SVMs), known for their ability to identify complex patterns in data.

Data pre-processing is crucial for ensuring the quality of our models. This stage involves cleaning inconsistencies, handling missing values, and potentially creating new features through feature engineering. A rigorous evaluation process using metrics like accuracy, precision, and recall will be employed to assess the performance of each model.

This multi-disease prediction system presents a unique opportunity for early disease detection, made accessible by Streamlit. Streamlit is a powerful Python library that allows us to design a user-friendly web application. Users can interact with the system through a simple interface, providing their data and receiving real-time predictions. This empowers individuals to take a more proactive approach to their health, potentially prompting them to seek further medical evaluation.

**Chapter 2**

**Literature review**

The ever-growing burden of chronic diseases like heart disease and diabetes necessitates advancements in early detection and preventative measures. Machine learning (ML) has emerged as a powerful tool for analysing vast amounts of medical data, offering promising avenues for multi-disease prediction systems. This literature review explores existing research on the application of ML for predicting heart disease and diabetes, laying the groundwork for our proposed system.

**Diabetes Prediction:**

Al Rashidi et al. [1] "Deep Learning Approaches for Diabetes Prediction: A Review": This review paper demonstrated the use of deep learning methods for diabetes prediction, which includes convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The authors focussed on the potential of these methods for extracting intricate patterns from a variety of data sources.

Chatterjee et al. [2] created a diabetes prediction model. In this paper they used a combination of machine learning methods such as logistic regression, decision trees, random forests, and support vector machines They used a dataset including clinical and demographic information, and it was able to predict the beginning of diabetes with an accuracy of 76%.

**Heart Disease Prediction:**

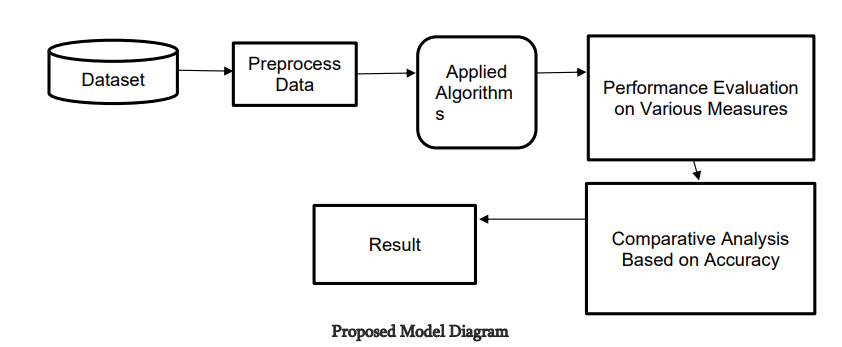
Chaimaa Boukhatem et al. [3] “Heart Disease Prediction Using Machine Learning”: This paper demonstrated the use of various Machine Learning algorithms for heart disease prediction such as SVM, Logistic Regression and Naïve Bayes. They had calculated the efficiency of these algorithms for heart disease prediction among which SVM has highest efficiency which was 76.8.

SenthilKumar Mohan et al. [4]” Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”: This paper demonstrated the hybrid model for heart disease prediction which uses hybrid Random Forest with linear model named as HRFLM. The efficiency of their model in this approach is almost 85%.

Aakash Chauhan et al. (2018) presented “Heart Disease Prediction using Evolutionary Rule Learning”. This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient’s dataset. This will facilitate (help) in decreasing the number of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness.

**Chapter 3 Proposed Model**

The approach employs Support Vector Machines (SVM) for diabetes prediction and Logistic Regression for heart disease prediction, which must undergo the following procedures:



**Figure 3(a) Proposed Model Diagram**

**Dataset:** We have collected the dataset for our project from reliable sources such as Kaggle which is famous for its data accuracy. We have collected the diabetes set which has 768 instances and 8 attributes. The attributes are as follows:

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: Diabetes pedigree function
* Age: Age (years)

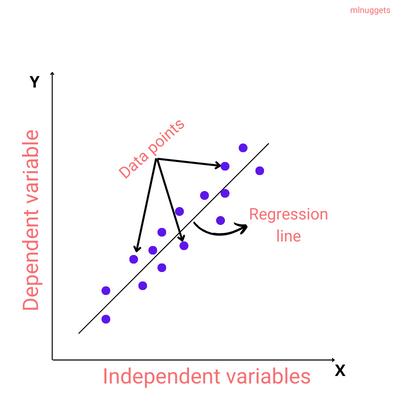
For Heart disease prediction the data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

The attributes of heart disease dataset are as follow:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) coloured by fluoroscopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect  
     .

**Algorithms:** We have worked on several algorithms such as linear regression, logistic regression, Support vector machine, K nearest neighbour and Naïve bayes.

**Linear Regression:** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, with the goal of finding the best-fitting line (or hyperplane) that minimizes the sum of squared differences between the observed and predicted values. This line can then be used to make predictions about the dependent variable based on new values of the independent variable. Linear regression is widely used in various fields, including economics, finance, and machine learning, for tasks such as prediction, forecasting, and understanding the relationship between variables.



**Figure 3(b) linear Regression diagram**

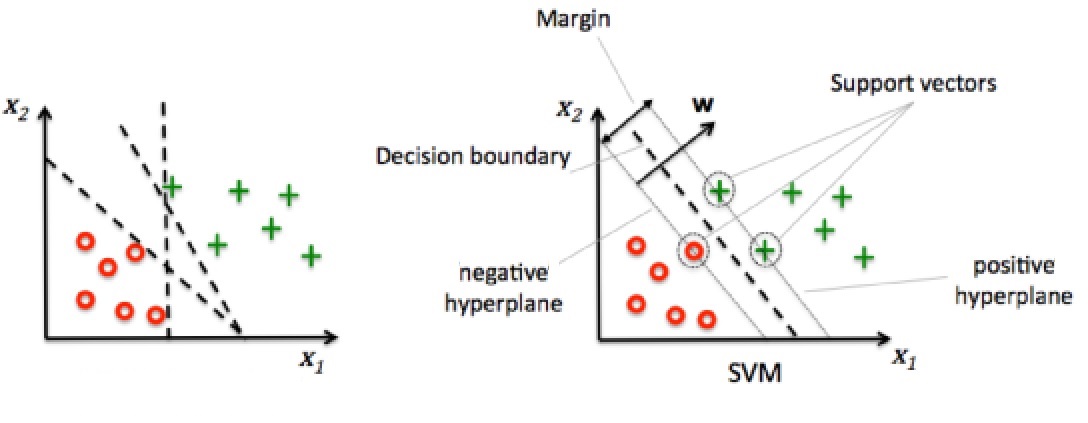
**Logistic Regression:** Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that an instance belongs to a particular class. Unlike linear regression, which predicts continuous values, logistic regression models the probability of the outcome using a logistic function. It estimates the odds of the event occurring by fitting a logistic curve to the data. The logistic function maps any input into a value between 0 and 1, representing the probability of the positive class. Logistic regression learns the coefficients of the independent variables to maximize the likelihood of the observed data. It's widely used in various fields, including healthcare, marketing, and finance, for tasks such as predicting customer churn, spam detection, and medical diagnosis.

A diagram of a curve

Description automatically generated

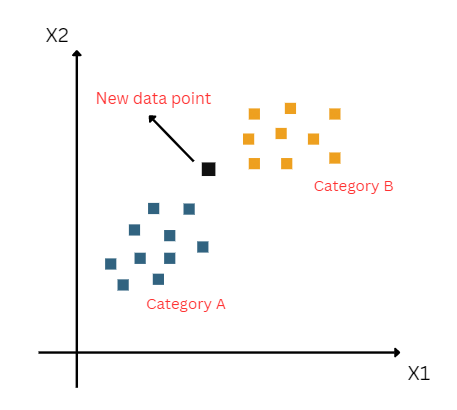
**Figure 3(c) Logistic Regression graph**

**Support Vector Machine:** Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates the data points into different classes. This hyperplane is chosen such that it maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, also known as support vectors. SVM can handle linear and non-linear classification tasks through the use of different kernel functions, such as linear, polynomial, radial basis function (RBF), and sigmoid. SVM is robust to overfitting, particularly in high-dimensional spaces, and is effective for datasets with clear margin of separation between classes. However, it may not perform well on large datasets due to its computational complexity.



**Figure 3(d) Support Vector Machine Diagram**

**K nearest Neighbour (KNN):** K-Nearest Neighbors (KNN) is a versatile and intuitive machine learning algorithm used for classification and regression tasks. It works by finding the k-nearest data points in the training set to the given test data point based on a chosen distance metric, commonly Euclidean distance. In classification, the majority class among the k-nearest neighbors is assigned to the test point, while in regression, the average or weighted average of the target values of the neighbors is predicted. KNN's simplicity makes it easy to implement and understand, but it can be computationally expensive, especially with large datasets. Additionally, selecting the appropriate value of k and the right distance metric are crucial for its effectiveness. Despite its limitations, KNN is widely used in various fields, including recommendation systems, image recognition, and anomaly detection.



**Figure 3(e) K Nearest Neighbour Graph**

**Naïve Bayes:** Naive Bayes is a popular and simple machine learning algorithm used for classification tasks. It's based on Bayes' theorem, which describes the probability of an event given prior knowledge. The "naive" assumption of Naive Bayes is that features are independent of each other, which simplifies the computation and makes it efficient even with large datasets.In classification, Naive Bayes calculates the probability of each class given the input features and then predicts the class with the highest probability. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature, hence the term "naive."

Despite its simplicity, Naive Bayes performs well in many real-world scenarios, especially when the independence assumption approximately holds true. It's particularly useful in text classification tasks, such as spam detection and sentiment analysis, where features are often word frequencies. Naive Bayes is also robust to irrelevant features and works well with high-dimensional data.

**3.1 Diabetes Prediction**

**Algorithm:** We have used Support Vector Machine (SVM) for diabetes prediction. SVMs offer a combination of robustness, flexibility, and interpretability that make them well-suited for diabetes prediction tasks, especially when working with medical datasets from sources like Kaggle.

**Data collection:** The data has been collected from Kaggle and UCI data set which contains attributes such as age, glucose, blood pressure, insulin, BMI etc.

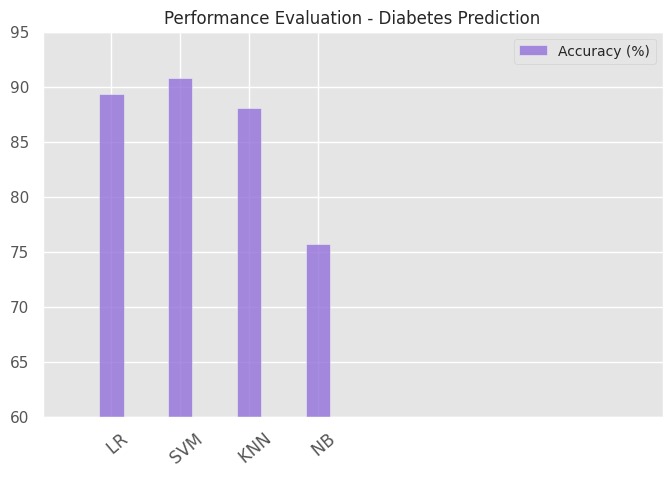
**Data Preprocessing:** Data preprocessing in diabetes prediction involves several essential steps to ensure the dataset's quality and suitability for machine learning models. This includes handling missing values through imputation or deletion, detecting and treating outliers, scaling numerical features to a similar range, encoding categorical variables into numerical representations, and addressing class imbalance if present.

**Feature Selection:** Feature selection is a crucial step in diabetes prediction to identify the most informative features that contribute to accurate model predictions. In this process, irrelevant or redundant features are removed, improving model efficiency, and reducing overfitting. Techniques such as correlation analysis, feature importance ranking, and domain knowledge are employed to select the most relevant features. These selected features, such as glucose levels, insulin sensitivity, BMI, age, and family history, provide valuable insights into diabetes risk factors. By focusing on the most impactful features, machine learning models can effectively distinguish between diabetic and non-diabetic individuals, leading to more accurate predictions in diabetes prediction tasks.

**Model Training:** Firstly, we have installed the required python libraries such as numpy, pandas, streamlit and sklearn. For diabetes prediction using Support Vector Machines (SVM), data is first prepared, featuring key parameters like glucose levels, BMI, Insulin, age etc . Model parameters are optimized through techniques like grid search. The SVM model is trained on the prepared data to discern diabetic from non-diabetic individuals. Evaluation metrics such as accuracy and F1-score gauge model performance. Hyperparameters are fine-tuned iteratively, and cross-validation ensures generalizability. The model's effectiveness is validated on unseen data. Finally, the trained SVM model is deployed for real-world prediction, aiding healthcare professionals and individuals in early detection and proactive management of diabetes.

**Comparison of Models for Diabetes Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.N.** | **Model Name** | **Train Accuracy(%)** | **Test Accuracy(%)** |
| **1** | Logistic regression | 88 | 87 |
| 2 | SVM | 91 | 90 |
| 3 | KNN | 87 | 86 |
| 4 | Naïve Bayes | 76 | 75 |



**Figure 3. (f) Performance evaluation of diabetes prediction**

**3.2 Heart Disease Prediction**

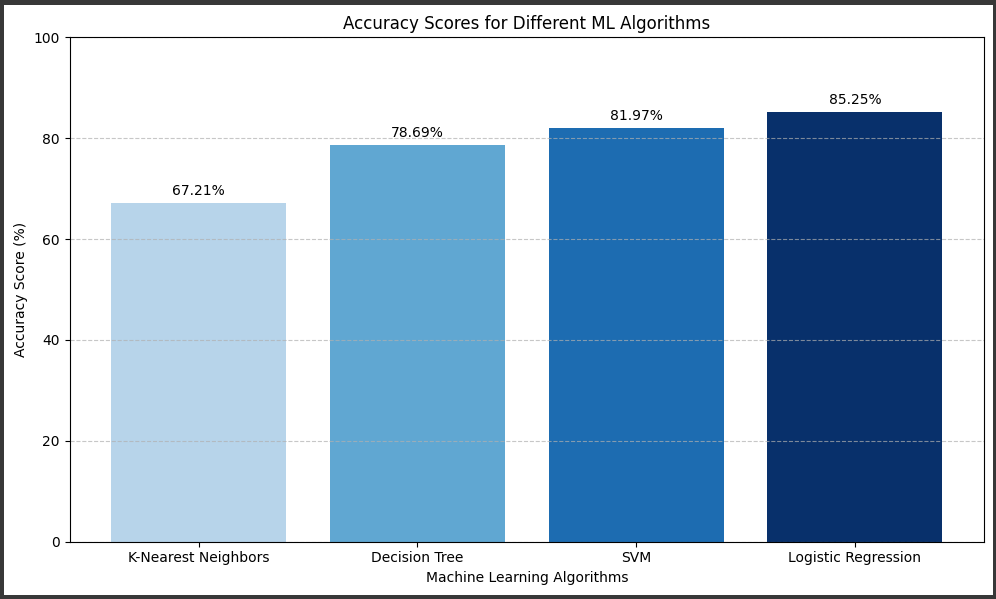
**Algorithm:** We have used Logistic Regression for heart disease prediction due to its simplicity and interpretability. It models the probability of heart disease occurrence based on input features such as age, sex, blood pressure, cholesterol levels, chest pain type etc . By fitting a logistic function to the data, it provides binary classification, distinguishing between individuals with and without heart disease, aiding in preventive healthcare measures.

**Data collection:** We have used Kaggle and UCI dataset for our data collection. The data collection contains attributes such as age, sex, chest pain type, blood pressure, cholesterol, blood sugar level on fasting etc.

**Data Preprocessing:** Data preprocessing for heart disease prediction using logistic regression involves steps to ensure data quality and model efficacy. This includes handling missing values, detecting and treating outliers, scaling numerical features for uniformity, and encoding categorical variables. Additionally, feature selection techniques such as correlation analysis are applied to identify relevant predictors. By preparing the dataset in this manner, logistic regression models can effectively distinguish between individuals with and without heart disease, facilitating accurate predictions and informed healthcare decisions.

**Model Training**: Model training for heart disease prediction using logistic regression involves several key steps. Firstly, the pre-processed dataset is split into training and testing sets. Then, logistic regression is applied to the training data, where the model learns the relationship between input features (such as age, blood pressure, cholesterol levels) and the likelihood of heart disease occurrence. The model's parameters are optimized using techniques like gradient descent to minimize the logistic loss function.

During training, the model iteratively adjusts its coefficients to maximize the likelihood of correctly predicting the presence or absence of heart disease. Cross-validation techniques may be employed to ensure the model's generalizability and prevent overfitting. The trained logistic regression model is then evaluated using the testing data to assess its predictive performance, typically using metrics such as accuracy, precision, recall, and F1-score. Finally, the validated model can be deployed for heart disease prediction in real-world scenarios, aiding healthcare professionals in early diagnosis and proactive management.



**Figure 3(g) Accuracy score for different machine learning algorithms**

**Comparison of Models for Heart Disease Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.N.** | **Model name** | **Train accuracy(%)** | **Test accuracy(%)** |
| **1** | Logistic Regression | **87** | **85** |
| **2** | SVM | **84** | **82** |
| **3** | KNN | **69** | **67** |
| **4** | Naïve Bayes | **84** | **81** |

**3.3 Streamlit Integration:** We have used Streamlit which builds a user-friendly interface for our disease prediction system Streamlit, a Python library, streamlines the development of data-driven web applications, simplifying the creation of interactive user interfaces for data visualization, exploration, and machine learning tasks. It offers an intuitive approach, sparing developers the complexities of writing HTML, CSS, or JavaScript code. With Streamlit, data scientists and developers can concentrate on the essential functionalities of their applications, enhancing productivity and efficiency. This library facilitates the seamless integration of various components, enabling users to interactively engage with data and machine learning models. By providing a straightforward workflow, Streamlit empowers users to quickly iterate on ideas, prototype applications, and deploy them effortlessly. Its user-friendly interface makes it accessible to a wide range of users, fostering collaboration and innovation in data science and application development.

Streamlit integration facilitates the creation of interactive web applications for machine learning models. For multiple disease prediction, leveraging logistic regression for heart disease and Support Vector Machine (SVM) for diabetes prediction offers robust solutions. Logistic regression models the probability of a binary outcome using a logistic function, making it suitable for heart disease prediction where the goal is to classify patients into diseased or healthy categories based on various risk factors. On the other hand, SVM aims to find the optimal hyperplane that separates classes in a high-dimensional feature space, making it ideal for diabetes prediction by distinguishing between diabetic and non-diabetic patients based on clinical features. With Streamlit, users can input their medical data interactively, obtain predictions, and visualize model performance metrics, providing a user-friendly interface for healthcare professionals and individuals seeking disease risk assessments. The integration of Streamlit ensures seamless deployment and accessibility of these predictive models to a wider audience, promoting informed decision-making and health management.

**Chapter 4**

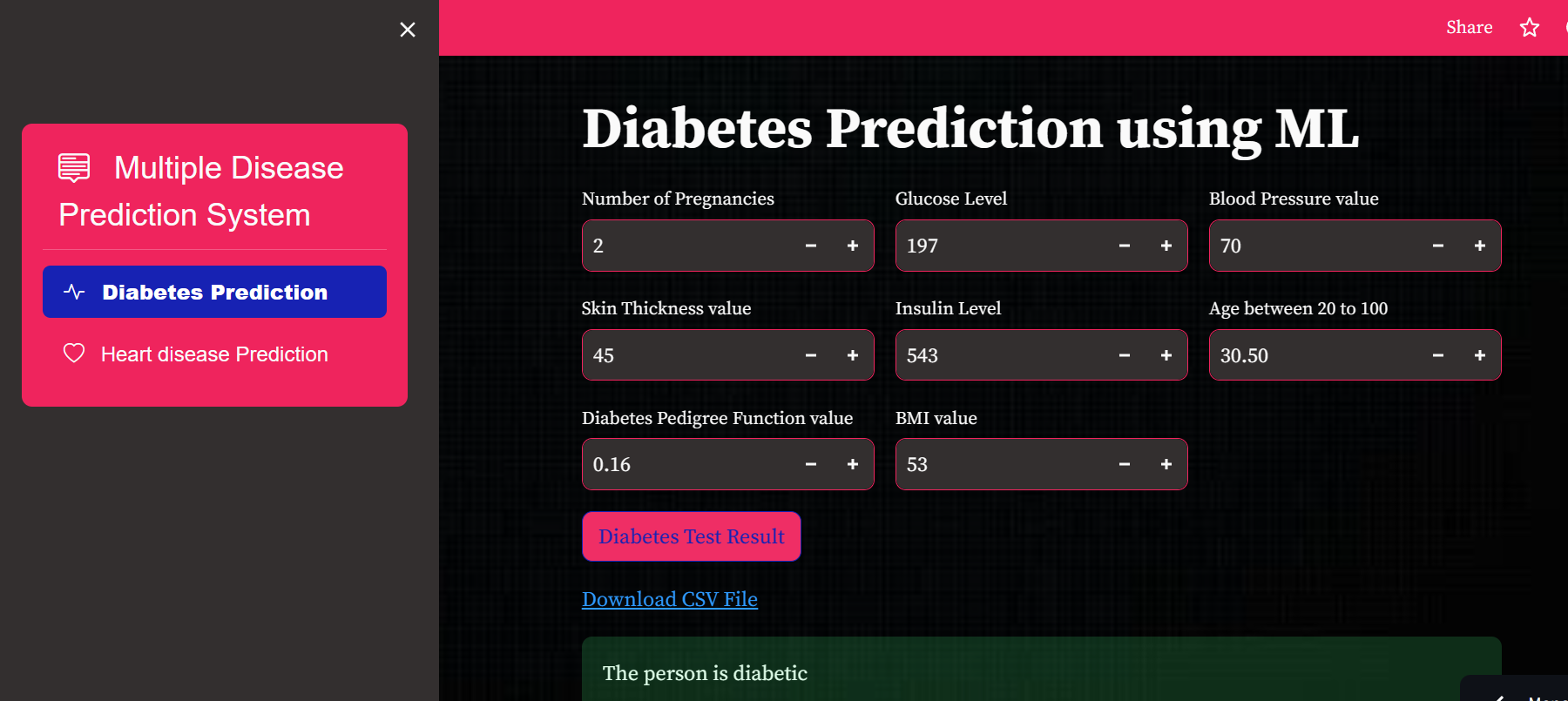
**Result**

In this project the Multiple Disease Prediction System powered by machine learning (ML) and facilitated through Streamlit's user interface, users can seamlessly predict the likelihood of both diabetes and heart disease. Through intuitive inputs provided via the Streamlit interface, users receive predictions generated by sophisticated algorithms. Support Vector Machines (SVM) analyse the input data to predict the probability of diabetes with accuracy of 90%, while Logistic Regression models assess the likelihood of heart disease based on the provided health parameters with an accuracy of 85%. These results are presented instantly, aiding users in making informed decisions about their health. Streamlit's user-friendly design enhances the overall experience, making predictive healthcare accessible and actionable for a wide range of users.

|  |  |  |
| --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** |
| Diabetes | Support vector machine | 90% |
| Heart disease | Logistic Regression | 85% |

**Diabetes Prediction System**

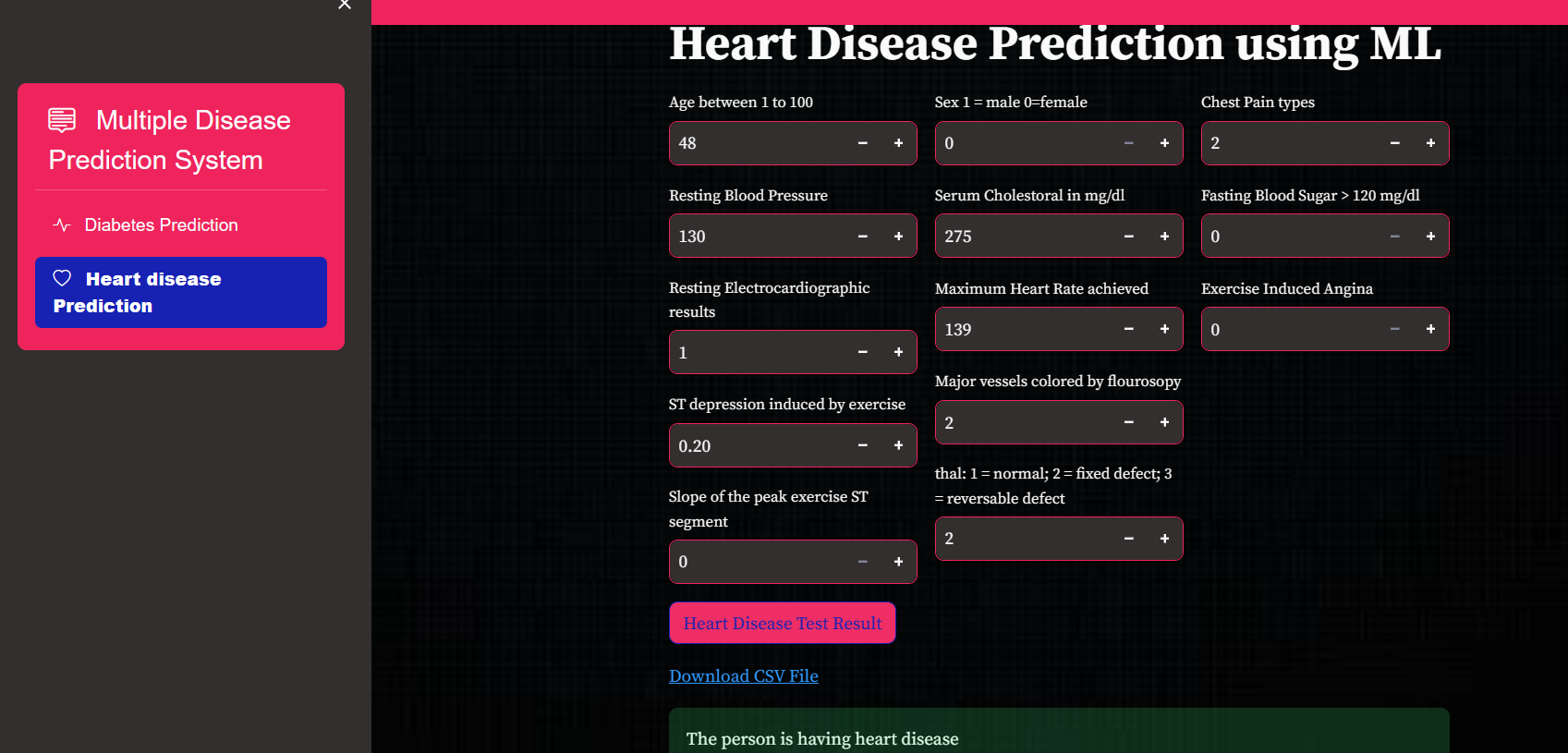




**Figure.4(a). Webpage of Diabetes Prediction System**

In fig 4(a) user gives the input for the given attributes such as Number of pregnancies, Glucose level, Blood pressure value, Skin thickness value, Insulin value, age, Diabetes pedigree function value and BMI value. After user gives the input, the system performs SVM algorithms and shows the output.

**Heart Disease Prediction System**



**Figure.4(b). Webpage of Heart Disease Prediction System**

In fig 4(b) user gives input values such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate, exercise induced agina, st depression value, slope etc. After user gives the input, the system performs the Logistic Regression and then shows the output on the screen.

**Chapter 5**

**Conclusion**

In conclusion, the multiple disease prediction system employing machine learning models has demonstrated promising accuracy rates in diagnosing two prevalent health conditions: diabetes and heart disease. Leveraging Support Vector Machine (SVM) for diabetes prediction and Logistic Regression for heart disease prediction, the system achieves respective accuracies of 90% and 85%.

Utilizing Streamlit for the user interface (UI) enhances accessibility and user experience, allowing individuals to input their health parameters seamlessly and receive instant predictions regarding their likelihood of having diabetes or heart disease.

The integration of SVM for diabetes prediction capitalizes on its ability to delineate complex decision boundaries, effectively classifying individuals based on their health metrics. Meanwhile, Logistic Regression proves robust in discerning patterns within the data to predict the probability of heart disease occurrence.

By empowering users with insights into their health risks through intuitive UI interactions, this system contributes to proactive healthcare management. With further refinement and integration of additional disease prediction modules, such as for hypertension or cancer, the system holds potential to become a comprehensive tool for early disease detection and prevention, thereby fostering improved health outcomes in diverse populations.

While these accuracies are promising, it's important to remember that the system is for informational purposes only. Given user input, the system predicts the likelihood of these diseases, categorized as Person has disease or not based on the predicted probabilities. Further research and data collection can potentially improve the system's accuracy and expand its capabilities in the future.

**Chapter-6**

**Future Scope**

Future work for the multiple disease prediction application could involve several avenues for improvement and expansion. Firstly, enhancing the accuracy and robustness of the machine learning models for both heart disease and diabetes prediction would be crucial. This could entail gathering larger and more diverse datasets, employing advanced feature engineering techniques, and experimenting with different model architectures.

Secondly, incorporating additional features and risk factors into the predictive models could improve their predictive power. For example, for heart disease prediction, factors such as lifestyle choices, dietary habits, and family history could be considered. Similarly, for diabetes prediction, factors like physical activity levels, ethnicity, and medical history could be included.

Furthermore, integrating real-time data sources such as wearable devices or electronic health records (EHRs) could enable continuous monitoring and personalized predictions for users. This would require developing robust data pipelines and APIs to handle streaming data and update the models accordingly.

Additionally, expanding the application to include predictive analytics for other prevalent diseases, such as hypertension, cancer, or respiratory conditions, could provide users with a more comprehensive health assessment tool.

Lastly, incorporating interpretability techniques to explain the model predictions and building features for user feedback and model explainability could enhance user trust and adoption of the application. This could involve visualizations, explanations, and recommendations tailored to individual users' health profiles and risk factors.

**References**

**[1]** **Tarik A. Rashid, S.M.A., Abdullah, R.M**, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8.

[2] **M.S. Mohanty, Bali, S. Chatterjee, M. Sarma, and R. Puravankara** (2019). Diabot is a predictive medical chatbot that use ensemble learning. 8(2), 6334-6340

[3] **Chaimaa Boukhatem, Heba Yahia Youssef, Ali Bou Nassif**, 2022. Heart Disease Prediction Using Machine Learning DOI: 10.1109/ASET53988.2022.9734880

[4] **Senthil Kumar Mohan, Gautam Srivastav** “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, July 2019

[5] **Reddy, S. Nava Bharath, KV Narasimha Reddy, SN Tirumala Rao, and K. S. M. V. Kumar**. "Diabetes Prediction using Extreme Learning Machine: Application of Health Systems." In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 993-998. IEEE, 2023

[6**] Andrabi, Syed Abdul Basit, and Inderjeet Singh**. "A Comparative Study of Machine Learning Techniques for Diabetes Prediction." In 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA),2022 Sep 21

[7] **M. S. Amin, Y. K. Chiam, K. D. Varathan**, ‘‘Identification of significant features and data mining techniques in predicting heart disease,’’ Telematics Inform., vol. 36, pp. 82–93, Mar. 2019

[8] **C. Sowmiya and P. Sumitra**, ‘‘Analytical study of heart disease diagnosis using classification techniques,’’ in Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS), Mar. 2017

[9] **Sisodia, D., Singh, L., Sisodia, S**., 2014. Fast andAccurate Face Recognition Using SVM and DCT,in: Proceedings of the Second InternationalConference on Soft Computing for ProblemSolving (SocProS 2012), December 28-30, 2012

[10] **Chaitrali S. Dangare, Sulabha S. Apte**, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 888)Volume 47No.10, June 2012.

[11] **E. Char**, et al., "Ethical Considerations in AI-Driven Healthcare," 2020. M. Wild, et al., "Streamlit for Machine Learning: Creating Interactive Web Apps in Python," 2022

[12] **P. Krishnaiah**, et al., "Predictive Modeling for Multiple Diseases Using Machine Learning with Feature Engineering," 2015

[13] **Syed Matheen Pasha and Shilpa Ankalaki**, "Diabetes and Heart Disease Prediction Using Machine Learning Algorithms", International Journal of Emerging Trends in Engineering Research, vol. 08, no. 07, jul 2020.

[14] **Shuge Ouyang**, "Research of Heart Disease Prediction Based on Machine Learning", 2022

[15] **M. Manwal, T. Maithani, S. Mall, K. C. Purohit and T. Choudhury**, "Cardiovascular Disease and Diabetes Disease Prediction Using Machine Learning and Streamlit API," 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2023, pp. 543-549, doi: 10.1109/SMART59791.2023.10428464.

[16] **N. Nazir** et al., "Heart Disease Prediction Using Machine Learning," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 62-66, doi: 10.1109/IC3I59117.2023.10397848.

[17] **V. Sharma, S. Yadav and M. Gupta**, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.

[18] **K. A. Barry, Y. Manzali, O. Labaybi, F. Rachid and M. Elfar**, "Heart Disease Prediction and Classification Using Machine Learning Algorithms," 2023 7th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, 2023

[19] **A. Mangal and V. Jain**, "Performance analysis of machine learning models for prediction of diabetes," 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2022

[20] **L. V. R. Kumari, P. Shreya, M. Begum, T. P. Krishna and M. Prathibha**, "Machine Learning based Diabetes Detection", 2021 ICCES, pp. 1-5, 2021.

[21] **N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas**, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction", IEEE Access, vol. 9, pp. 103737-103757, 2021.

[22] **L. D. Gopisetti, S. K. L. Kummera, S. R. Pattamsetti, S. Kuna, N. Parsi and H. P. Kodali**, "Multiple Disease Prediction System using Machine Learning and Streamlit," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023

[23] **Rayan Alanazi**, "Identification and prediction of chronic diseases using ma chine learning approach", Journal of Healthcare Engineering, vol. 2022, 2022.