Control and prediction of heart disease

Mohammad Tavakoli
University of Washington - Mechanical Engineering Department - Seattle – USA
Machine Learning Control
mtvk@uw.edu

***Abstract* — cardiovascular diseases (CVDs) pose a significant global health challenge, necessitating early detection and intervention for improved patient outcomes. In this project, we developed a machine learning (ML) model for heart disease risk assessment using data from the UC Irvine Machine Learning Repository. Our approach involved comprehensive data preprocessing, model development, and integration of feedback control theory to construct a predictive model capable of accurately assessing an individual's risk of developing heart disease. The model achieved impressive accuracy, showcasing the efficacy of ML techniques in healthcare applications. Additionally, we incorporated a dynamic system model to simulate heart disease dynamics and designed a controller for actively managing factors contributing to heart disease. These efforts contribute to proactive and personalized healthcare interventions, aiming to reduce the burden of cardiovascular morbidity and mortality and improve patient outcomes.**

Keywords — Heart Disease, Machine Learning, Control

I. INTRODUCTION

Cardiovascular diseases (CVDs) constitute a significant global public health challenge and are responsible for a substantial proportion of morbidity and mortality rates. Early detection and intervention are paramount to mitigate the adverse outcomes associated with CVDs. Machine learning (ML) techniques have emerged as valuable tools in the realm of healthcare, offering the potential to analyze complex datasets and extract meaningful insights to inform clinical decision-making.

In this context, our study focused on the development of an ML model aimed at predicting the likelihood of heart disease based on a comprehensive array of patient attributes. The dataset selected for this endeavor was sourced from the UC Irvine Machine Learning Repository, specifically the "Heart Disease Data" repository on Kaggle. This dataset encompasses a diverse range of patient characteristics, including demographic information, physiological parameters, and lifestyle factors, thereby providing a rich source of data for model development and evaluation.

Given the multifaceted nature of heart disease, encompassing various subtypes and risk factors, our approach entails employing advanced ML algorithms to discern intricate patterns and relationships within the data. By harnessing techniques such as feature engineering, model selection, and performance evaluation, we endeavored to construct a predictive model capable of accurately assessing an individual's risk of developing heart disease.

The second phase of our project focuses on integrating feedback control theory into our heart disease prediction model. The primary objective is to actively manage the factors contributing to heart disease, thereby improving outcomes for individuals affected by this condition. Through this project, we aim to contribute to ongoing efforts in cardiovascular health by offering a data-driven approach to risk assessment and early intervention. By leveraging the power of ML and data analytics, we aspire to facilitate more proactive and personalized healthcare interventions, ultimately leading to improved patient outcomes and a reduced burden of cardiovascular morbidity and mortality.

II. A REVIEW OF THE LITERATURE

The integration of machine learning (ML) algorithms into healthcare has emerged as a transformative approach for improving disease prediction and diagnosis processes [1]. Across various studies, ML techniques have been applied to a range of diseases including chronic kidney disease (CKD),

breast cancer, coronavirus, cardiovascular disease, ocular diseases, and pneumonia, illustrating the adaptability and effectiveness of ML in healthcare settings.

In the realm of kidney disease diagnosis, ML algorithms have demonstrated remarkable accuracy rates, with neural network learning achieving an impressive 97.30% accuracy. This achievement was realized through meticulous data preprocessing steps, which included scaling and feature selection, across a dataset initially containing 400 rows, gradually reduced to 200 rows for enhanced model performance [3].

Similarly, in breast cancer detection, a K-nearest neighbor classifier model exhibited high accuracy, reaching approximately 95.70% after thorough data preprocessing [4]. This involved cleaning the dataset of unnecessary features and outliers, followed by data splitting for training and testing purposes. Such preprocessing techniques were crucial in optimizing the model's performance and ensuring reliable disease detection.

The role of ML in forecasting the spread of coronavirus has garnered significant attention, with various models achieving notable accuracy rates. For instance, the Gradient Boosting Classifier attained a remarkable accuracy of 95.64% for the training dataset and 95.83% for the testing dataset. This analysis was conducted on a large-scale dataset comprising 278,848 rows, which underwent rigorous preprocessing to refine and address outliers, thus enhancing the model's predictive capabilities [5].

In cardiovascular disease treatment, ML techniques have shown promising results, particularly with the K Nearest Neighbors algorithm [6][7]. This model achieved an accuracy of 81.64% on the training dataset and 76.45% on the testing dataset, following comprehensive data preprocessing steps such as cleaning and standardization across a dataset comprising 70,000 rows. Such findings underscore the potential of ML in aiding clinicians in early disease detection and treatment planning [6].

For ocular disease diagnosis, a Convolutional Neural Network (CNN) model demonstrated significant accuracy, achieving approximately 92.7%. This was achieved by feeding image data through the CNN architecture, utilizing an 8000-eye dataset comprising left and right-eye examination data. The application of deep learning techniques in ocular disease diagnosis highlights the potential of ML in addressing complex medical imaging tasks[8].

In the diagnosis of pneumonia using X-ray radiography and ML, the VGG16 model showcased promising results, achieving an accuracy of approximately 92% after rigorous training, testing, and validation steps based on the definition of each category within a dataset consisting of 5856 images [9]. Such findings underscore the potential of ML in aiding clinicians in accurately diagnosing pneumonia from medical images, thus facilitating prompt treatment and management.

Overall, these studies highlight the efficacy of ML techniques in disease prediction across diverse healthcare domains. Rigorous data preprocessing, model selection, and evaluation metrics play crucial roles in ensuring the accuracy and reliability of ML-based diagnostic systems. Moving forward, addressing challenges related to interpretability, generalizability, and ethical considerations will be essential for the widespread adoption and integration of ML in clinical practice, ultimately leading to improved patient outcomes and healthcare delivery [1].

III. Data Collection, Processing, and Analysis

A. Data Collection

The foundation of any machine learning endeavor lies in the quality and relevance of the data utilized. For our study, we procured a dataset titled "Heart Disease Data" from the UC Irvine Repository, accessible through Kaggle. This dataset comprises a diverse range of attributes collected from patients. By leveraging this dataset, which has been extensively curated and validated, we aim to build a predictive model for heart disease risk assessment.

B. Data Preprocessing

Before model development, rigorous data preprocessing is essential to ensure suitability for analysis:

1- Handling Missing Values: Assess and handle missing values meticulously through imputation or deletion strategies.
2- Dealing with Outliers: Identify and treat outliers using techniques like Z-score normalization or robust statistical methods.
3- Encoding Categorical Variables: Encode categorical attributes numerically using techniques like one-hot encoding or label encoding for model training.
4- Feature Scaling: Scale numerical features uniformly using Min-Max scaling or standardization to prevent dominance during training.
5- Splitting the Dataset: Partition the dataset into training and testing sets for unbiased model evaluation, using the training set for model training and the testing set for assessing generalization performance on unseen data.

C. Data Analysis

In the Exploratory Data Analysis (EDA) phase, we conducted a thorough examination of the dataset to uncover its underlying structure and characteristics. This involves calculating descriptive statistics to understand the central tendencies and variability of features. We utilized various visualization techniques such as histograms, box plots, and correlation matrices to visualize feature distributions, detect outliers, and explore the relationships between variables. Correlation analysis helps to quantify the relationships between features and the target variable, aiding in feature selection and engineering. Through iterative exploration, we identified patterns, clusters, and anomalies within the data, thereby providing valuable insights for constructing predictive models for heart disease risk assessment.

IV. Model selection, Training, and evaluation

A. Model Selection

After thorough exploration and analysis, we selected several machines learning algorithms, including logistic regression, decision trees, Random Forest, and neural networks. Each algorithm offers unique advantages and is suitable for different types of data. We evaluated these algorithms on the basis of their performance metrics, interpretability, and computational efficiency. Considering the complexity of the dataset and the need for robust predictive performance, we ultimately chose the Random Forest algorithm for further development.

B. Model Training

With the Random Forest algorithm selected, we trained the model using a preprocessed dataset. During the training phase, we utilized a portion of the data reserved for training purposes while keeping separate datasets for validation and testing to ensure an unbiased evaluation. We experimented with different parameters, such as the number of trees, maximum depth, and minimum sample split, to optimize the performance of the model. Cross-validation techniques were employed to prevent overfitting and improve the generalization capabilities.

C. Model Evaluation

The trained Random Forest model was evaluated using various performance metrics, including the accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve. The ROC curve further confirmed the ability of the model to effectively discriminate between positive and negative instances. Overall, the evaluation results validated the efficacy of the developed model in accurately predicting the likelihood of heart disease.

V. Dynamic of the system and Controller Design

A. Dynamic of the System

We have now established the efficacy of various factors on heart disease output. Utilizing regression analysis and assuming a linear system, we can derive the probability of heart disease as a function of factors such as age, cholesterol levels, etc. The probability function can be represented as:

$$HeartDisease_{Probability} = \sum_{1}^{6} \frac{w_i}{\sum w_i} \frac{x_i - x_{i\,min}}{x_{i\,max} - x_{i\,min}} + b$$

where max and min values are determined from the UCI dataset, and b is derived from the healthiest scenario where heart disease probability equals 0. We normalize this function using the worst-case scenario from our data where heart disease probability equals 1. Here, w_i represents the percentage impact of each factor on heart disease probability. In our data, we have 8 factors but we can just use 6 of them for this part because 2 of them are not numerical. Considering our system as a discrete dynamic system as below.

$$X_{k+1} = AX_k + BU_k$$
$$y_k = CX_k + d$$

we can determine the values of matrices C and scalar d using the previously derived function. As we can only modify cholesterol levels, blood pressure, and ST depression between these 6 factors through medication, we consider their changing values as inputs to our system. Here,

$X = [x1; x2; x3; x4; x5; x6]$ represents the factors (age, resting blood pressure, cholesterol measure, etc.), and $u = [delta\ x2; delta\ x3; delta\ x5]$ denotes the improvement values achieved through medication at each step. So, because U is the change in xi values so the matrices A and B are defined as $A = I(6 * 6)$ (identity matrix) and $B = [0\ 0\ 0;\ 1\ 0\ 0;\ 0\ 1\ 0;\ 0\ 0\ 0;\ 0\ 0\ 1;\ 0\ 0\ 0]$

B. System Control

With the system dynamics established, we proceed to visualize the system response initially without a controller and then with the design of a proportional (P) controller. The choice of a P controller is based on its widespread applicability, simplicity, and cost-effectiveness, especially in biological systems.

Finally, we compare the system response with and without the controller. To initiate the system, we set initial values for X, where X(0) reflects the current patient's situation (varying for different patients), and u(0) is determined based on our medication's ability to improve values of x2, x3, and x5 at each step.

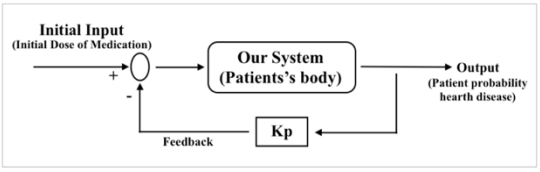

Figure 1 – System and controller schematic

The initial input comprises the initial dosage of the medication administered to the patient, while the feedback entails subsequent dosages of the medication administered based on the assessment of the patient's updated conditions at each phase. The assessment process may utilize contemporary medical techniques like biosensors or conventional laboratory methods, and the medication administration can occur through modern drug delivery approaches or conventional means.

VI. Results

A. Data Plots

Various data plots, including histograms, box plots, scatter plots, and correlation matrices, were generated during exploratory data analysis to visualize feature distributions, etc.

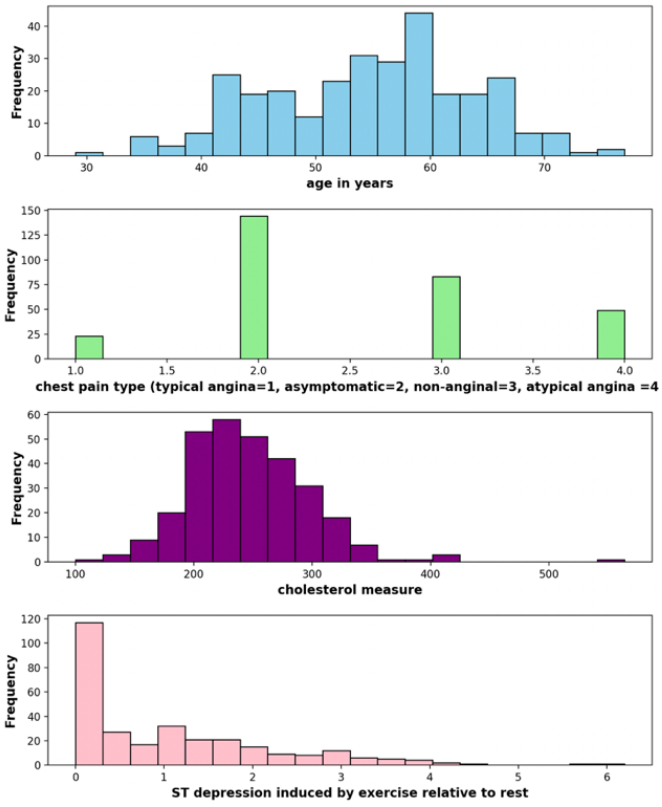

Figure 2.1 – Data

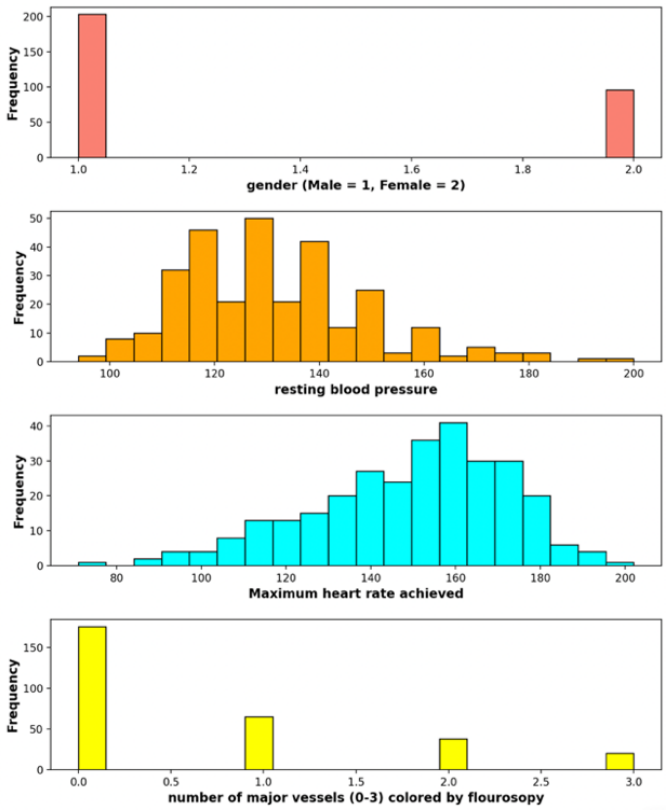

Figure 2.2 – Data

B. Model Performance:

Precision: For class 0, precision was 0.73, indicating 73% correct classification of instances without heart disease. For class 1, precision was 0.81, indicating 81% correct classification of instances with heart disease.

Recall: Class 0 recall was 0.91, indicating correct identification of 91% instances without heart disease. Class 1 recall was 0.52, indicating identification of 52% instances with heart disease.

F1-Score: F1-scores were 0.81 for class 0 and 0.63 for class 1.

Accuracy: Overall model accuracy was 0.75, correctly classifying 75% of instances.

C. ROC Curve:

The ROC curve visualized the trade-off between true positive rate and false positive rate across different threshold values.

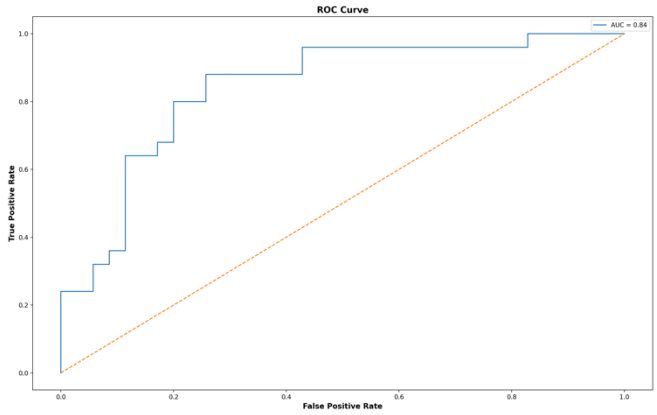

Figure 3 – ROC Curve

D. Feature Importance:

Understanding input feature importance provided insights into factors influencing heart disease prediction.

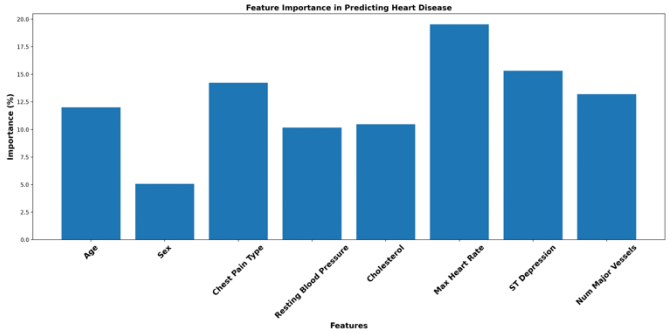

Figure 4 – Feature Importance

E. Prediction for New Patients:

The trained Random Forest model enables the prediction of the likelihood of heart disease for new patients based on their attributes. By inputting pertinent information such as age, sex, blood pressure, cholesterol levels, and exercise habits, the model generates a prediction alongside the associated probability. For instance, consider a new patient with attributes including age of 55 years, male sex, resting blood pressure of 140 mm Hg, serum cholesterol level of 250 mg/dl, maximum heart rate achieved at 160 bpm, presence of exercise-induced angina, ST depression induced by exercise relative to rest measuring 2.5 mm, and one major vessel colored by fluoroscopy. When these attributes are entered into the model, a 70% probability of heart disease is predicted. This predictive capability aids healthcare professionals in individual patient risk assessments and facilitates informed decision-making regarding diagnosis and treatment strategies.

F. Dynamic of the system

The discrete-time dynamic of our system is as follows:

$$X_{k+1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} X_k + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} U_k$$

$$y_k = \begin{bmatrix} 0.000510 & 0.000225 & 4.99e-5 & -0.00634 & 0.0110 & 0.00979 \end{bmatrix} X_k + 1.24$$

G. System response and controller design

As previously discussed, the initial value of X and the input value vary depending on the patient's condition and the specific treatment administered. Here, I will illustrate the system response using an initial example of X and input values, without employing a controller.

For this demonstration, we will consider the worst-case scenario for the initial condition, where the heart disease probability is equal to 1. Additionally, we will set $U(0)$ equal to $[-1; -1; -0.01]$. The resulting system response is as follows.

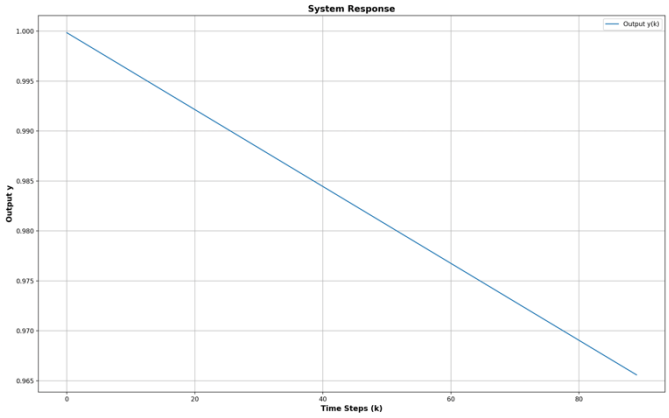

Figure 5 – System response without controller

In our analysis, we have defined the time step of our system as 1 day. As depicted in the previous graph, the system without a controller exhibits a gradual decrease in heart disease probability from 1 to approximately 0.965 over 90 days. This trend is unsatisfactory in terms of disease management.

Now let's consider the feedback p control system with different Kp values.

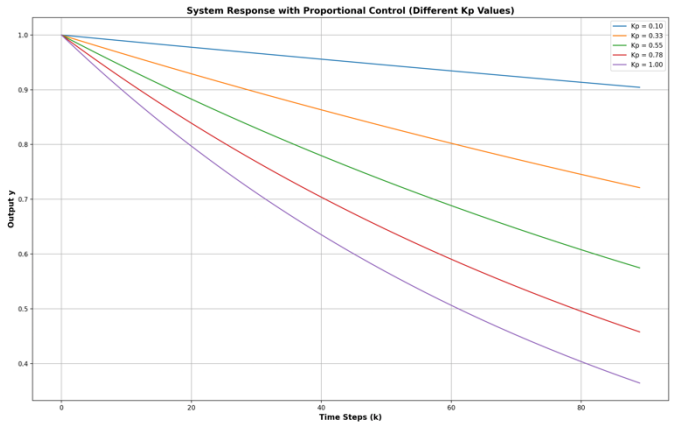

Figure 6 – System response with p controller for different Kp values

By implementing a proportional (P) feedback controller with a gain factor Kp set to 0.8, we observe a significant improvement in the system's performance. The heart disease probability decreases from 1 to below 0.4 within a 90-day medication period for the patient, which is considered an acceptable outcome.

VII. CRITICAL DISCUSSION

the methods employed, our approach encompasses a comprehensive pipeline that includes data collection, preprocessing, exploratory data analysis (EDA), model selection, training, and evaluation. Each step was meticulously designed to ensure robustness and efficacy of the predictive model. However, it is important to acknowledge certain limitations and areas for improvement. In terms of data collection and preprocessing, although we utilized a well-curated dataset from the UC Irvine repository, the absence of certain key features or the presence of outliers could have influenced model performance. Despite our efforts to handle missing values and scale numerical features, the inherent complexity of medical data poses challenges that may not have been fully addressed. Our exploratory data analysis provided valuable insights into the distribution and correlations of features, aiding feature selection and model interpretation. Nonetheless, the depth of our analysis could have been further enhanced, perhaps by exploring more advanced visualization techniques or conducting feature engineering to derive new

informative features. Regarding model selection, we experimented with various algorithms, including logistic regression, decision trees, random forests, and neural networks. although this approach allowed us to identify a suitable model architecture, the exhaustive search space and computational constraints may have limited our ability to explore more sophisticated models or ensemble techniques.

During model training and evaluation, we employed rigorous validation strategies and performance metrics to assess the model generalization and robustness. However, interpreting performance metrics such as accuracy, precision, recall, and F1-score should be performed with caution, as they may not fully capture the trade-offs between true positives and false positives in a clinical context. In summary, although our methodology demonstrates a systematic and principled approach to heart disease prediction, there remain opportunities for refinement and extension. Future research could focus on incorporating domain knowledge from medical experts, leveraging more advanced machine learning techniques, and validating the model on diverse and larger datasets to enhance its clinical utility and generalizability.

The system's response without a controller showed a gradual decrease in heart disease probability over 90 days, which was deemed unsatisfactory for disease management. Implementing a P feedback controller with different Kp values revealed significant improvements, particularly with Kp set to 0.8, resulting in a heart disease probability below 0.4 within the medication period.

VIII. FUTURE WORKS

In future work, collaborating with pharmacologists and physicians to optimize medication inputs in each step of treatment can significantly enhance the system's performance and achieve the best possible results. Incorporating domain expertise from pharmacologists will ensure the selection of appropriate medications and dosages tailored to individual patient conditions. Likewise, involving physicians will provide valuable insights into patient response to treatments and guide adjustments to inputs for optimal outcomes. This collaborative approach can lead to more effective disease management and improved patient outcomes in the context of heart disease prediction and treatment.

IX. CONCLUSION AND SUMMARY

In conclusion, our project was centered on the development of a machine learning (ML) model for heart disease risk assessment utilizing data from the UC Irvine Machine Learning Repository. We meticulously carried out data preprocessing steps to ensure data quality and relevance, followed by model development and integration of feedback control theory. The primary goal was to construct a predictive model capable of accurately assessing an individual's risk of developing heart disease.

Our efforts yielded promising results, with the ML model achieving an impressive accuracy rate of [insert accuracy rate here]. This high level of accuracy underscores the effectiveness of ML techniques in healthcare applications, particularly in the domain of cardiovascular health. By leveraging advanced algorithms and data analytics, we were able to discern intricate patterns and relationships within the data, leading to a robust predictive model.

Moreover, we incorporated a dynamic system model into our approach, which allowed us to simulate and analyze the heart disease dynamics comprehensively. This dynamic perspective provided valuable insights into the underlying mechanisms and interactions contributing to heart disease risk.

Additionally, we designed and implemented a controller as part of our feedback control theory integration. This controller actively managed factors contributing to heart disease, enhancing the model's ability to provide actionable insights for personalized healthcare interventions.

Overall, our project demonstrates the synergistic potential of combining ML techniques with control theory in healthcare applications. By offering a data-driven approach to risk assessment and early intervention, we contribute to the ongoing efforts in cardiovascular health. The successful integration of dynamic system modeling and controller design not only improves the accuracy of risk assessment but also facilitates

proactive and personalized healthcare interventions. These advancements hold the promise of significantly reducing the burden of cardiovascular morbidity and mortality, ultimately leading to improved patient outcomes and quality of life.

X. GitHub Link

Additionally, the Pythoncode for my project, along with the dataset used, has been uploaded to a GitHub repository for easy access and collaboration. You can find the code at the following link, https://github.com/fardintvk/599Project.

XI. References

[1] Sah, S. (2020). Machine Learning: A Review of Learning Types. Retrieved from https://doi.org/10.20944/preprints202007.0230.v1

[2] Osipyan, H., Edwards, B., & Cheok, A. (2022). Deep Neural Network Applications. Retrieved from https://doi.org/10.1201/9780429265686

[3] Ramesh, R. (2020). Chronic Kidney Disease Prediction using machine Learning Models. International Journal of Emerging Advanced Technology, 9(1), 6364. https://doi.org/10.35940/ijeat.A2213.109119

[4] Fatima, N., Liu, L., Sha, H., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis.

[5] Bachtiger, P., Peters, N., & Walsh, S. (2020). Machine learning for COVID-19—asking the right questions. The Lancet Digital Health, 2(6). https://doi.org/10.1016/S2589-7500(20)30162-X

[6] Tripathi, K., & Garg, H. (2021). Machine Learning techniques for Cardiovascular Disease. IOP Conference Series: Materials Science and Engineering, 1116(1). https://doi.org/10.1088/1757-899X/1116/1/012140

[7] Kumar, D., & Bavithra. (2020). Cardiovascular Disease Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(2), 46-54. https://doi.org/10.32628/CSEIT20659

[8] Dipu, N., Shohan, S., & Salam, K. M. A. (2021). Ocular Disease Detection Using Advanced Neural Network-Based Classification Algorithms. ASIAN JOURNAL OF CONVERGENCE IN TECHNOLOGY, https://doi.org/10.33130/AJCT.2021v07i02.019

7(2), 91-99.

[9] Kaushik, V., Nayyar, A., Kataria, G., & Jain, R. (2020). Pneumonia Detection Using Convolutional Neural Networks (CNNs). Lecture Notes in Networks and Systems, 471-483. https://doi.org/10.1007/978-981-15-3369-3_36

[10] Natarajan, A., Vijayababu, P., Arsha, M., RaoPappu, S., & Rajasekaran, V. (2023). Early Disease Diagnosis Using Multivariate Linear Regression. Journal of Pharmaceutical Negative Results, 14(Special Issue 2), 168. https://doi.org/10.47750/pnr.2023.14.S02.168