CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1. INTRODUCTION

The Part 2 report is a supplement for the Part 1 Report submitted earlier titled "A Fuzzy Logic System for Diagnosing if an Individual has / is at Risk of Heart Disease" Given its growing popularity and implementation in the medical field, the study aims to design and build an automated and simple fuzzy logic system to predict the level of risk an individual is to heart disease. This was done to simplify the process of diagnosis whilst still be able to predict with high accuracy and to try and mitigate the shortcomings traditional methods of diagnosis. By doing so, initiatives can be taken by individuals and medical providers to curb or reduce the risk of exposure to the disease. Part 1 of this project provided a brief overview to the field of Artificial Intelligence and the topic of Heart disease. Subsequently, the problem statement and aim of the study was defined. A detailed literature review was conducted to identify several inputs that are risk factors / indicators used to commonly diagnose if an individual is at risk of heart disease. The fuzzy linguistic relationship of each input and the output which is the category of risk for heart disease were then defined. Five inputs were determined, and they are chest pain, cholesterol, systolic blood pressure, blood sugar and age.

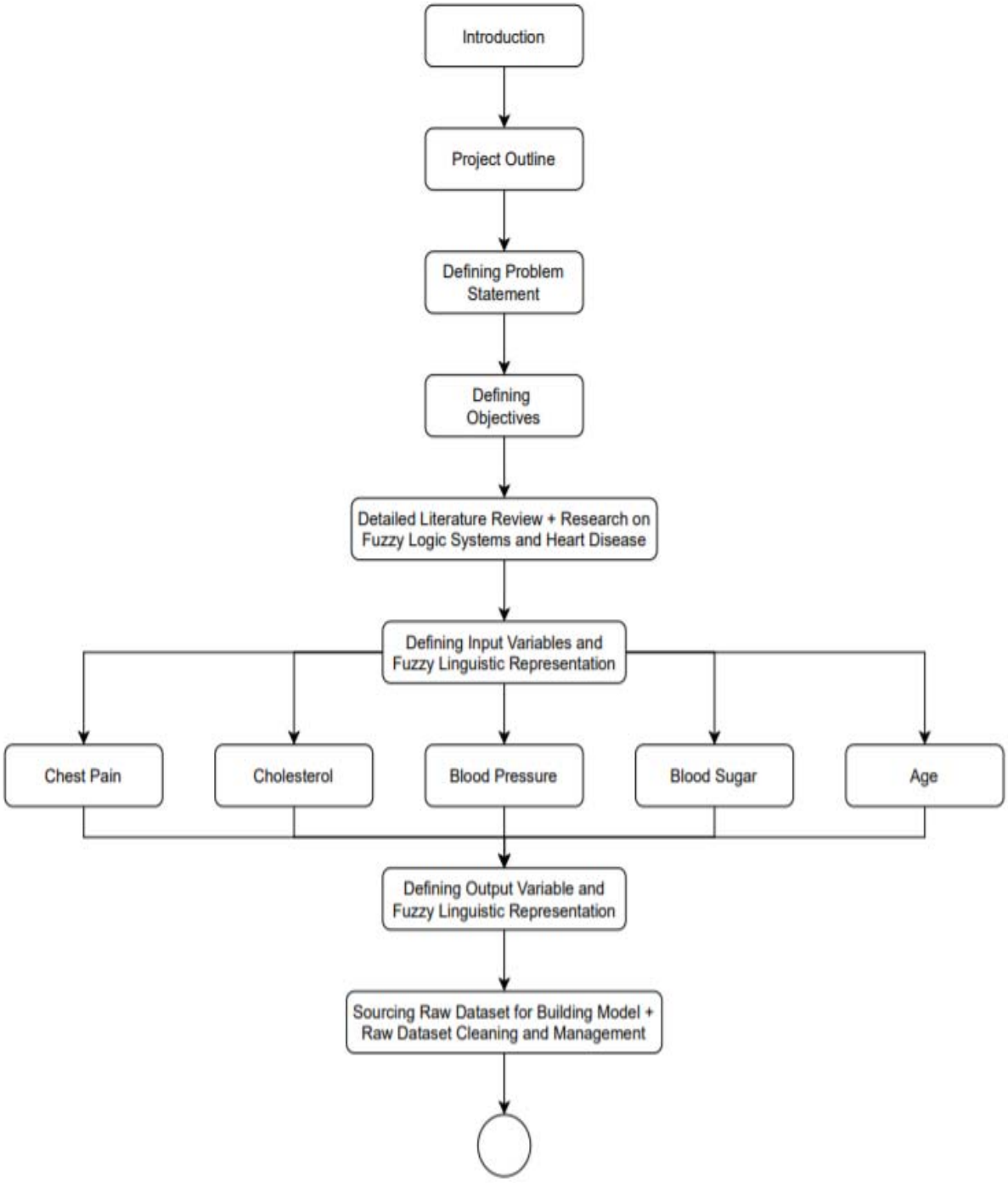
1) Chest pain was categorized into 1 - Typical Angina, 2 - Atypical angina, 3 - Non-angina pain and 4- Asymptomatic.
2) Serum Cholesterol level was categorized into Low (<200 mg/dL), Medium (194-245 mg/dL) and high (>240 mg/dL).
3) Systolic Blood Pressure was categorized into Low BP (<125), Medium BP (120-140) and High BP (>130)
4) Blood sugar was categorized into a binary Fuzzy Class of Low Blood Sugar (<120) and High Blood Sugar (≥240)
5) Age was segregated into Young (<33), Middle (33-45) and Old (>45).

Based on the inputs defined in Part 1, Part 2 will detail the implementation of the Fuzzy Logic Mamdani Model for diagnosis of heart disease. The chronology of the report is as follows:

1) Explanation on the characteristics of the raw dataset used for this study.
2) Defining the membership function selection for each input parameter as well as the final output and the final selected membership function for all parameters. Along with justification.
3) Defining the rules for the Mamdani model and final selection of rules to be aggregated in the final model along with justification.
4) Justification of Fuzzy Logic System Model Chosen i.e., Mamdani model.
5) Building the final model and handling exceptions.
6) Discussion on the results and findings.
7) Conclusion and final recommendations.

2. METHODOLOGY

Figure 1 below provides a flowchart of the methodology executed for this project.



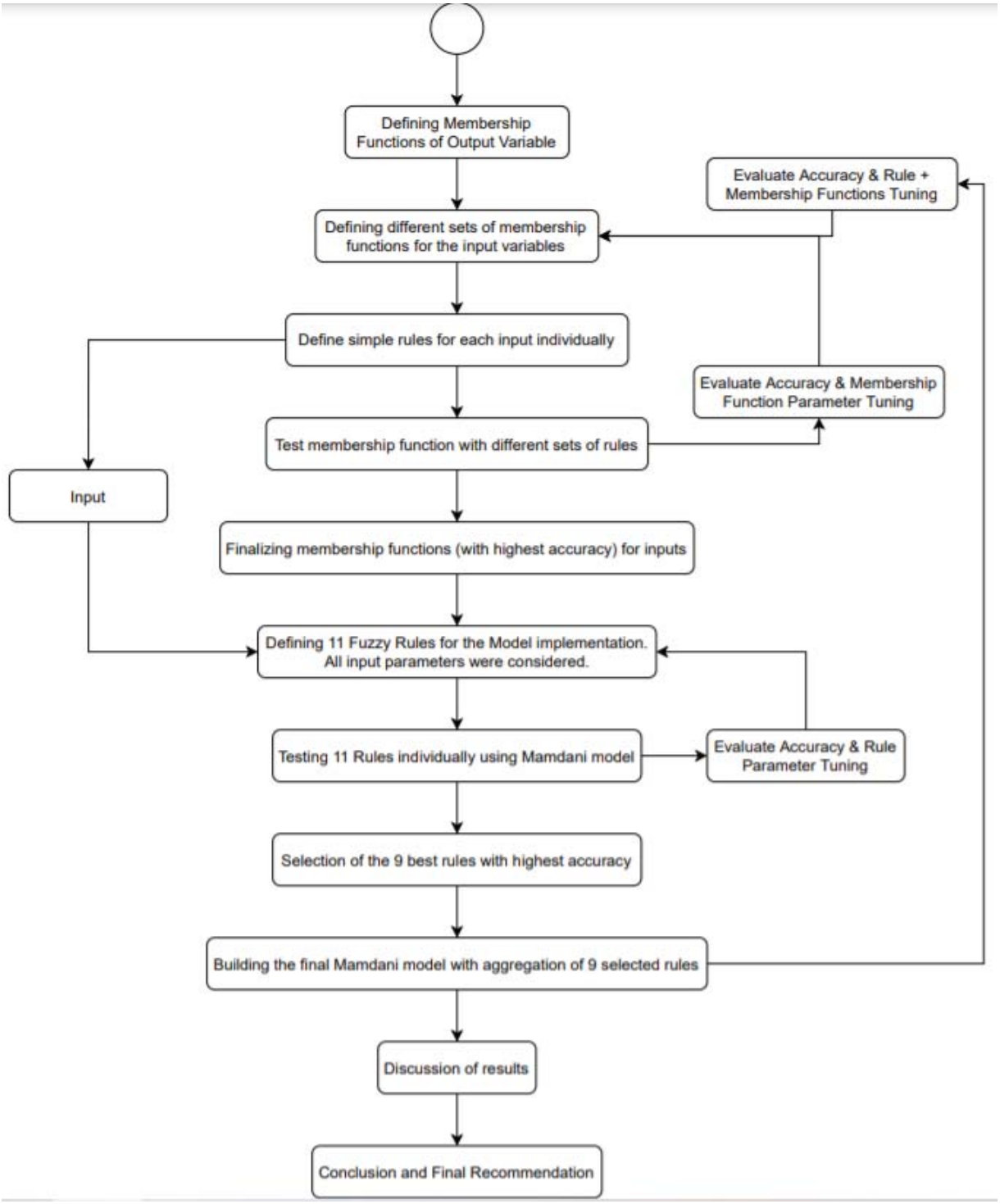


Figure 1: Process Flowchart

3. PROGRAMMING LANGUAGE FOR IMPLEMENTATION

The project was executed entirely in Python programming in the PyCharm Integrated Development Environment (IDE). Python language was selected as the program to execute this project because of its simplicity and consistency, its abundance of great libraries and frameworks for AI, data processing, data visualization and Machine Learning (ML). It also has flexibility, platform independence and because it is widely used for these sorts of applications making it a very practical and justifiable language to implement. PyCharm was used as it is one of the most commonly available and free IDE available. Several packages were installed and imported into the implementation script during the project execution. The packages and their functions are listed below in Table 1.

Name of Python Package (abbreviation)	**Function**
Pandas (pd)	- For raw data importing from the project repository - For managing, rearranging the data into a desirable form, selecting data features that were only desired for the study and to remove data rows with missing values.
Memberships (mm)	- Project specific.py file containing various membership functions created for implementation in inputs' individual .py files and the final model
Numpy (np)	- NumPy library was imported for creating, working, and carrying out computational operations with arrays and matrices.
Pyplot from Matplotlib (plt)	- Pyplot was imported for plotting the and visualizing the membership functions, the raw data and results.
Confusion_matrix from sklearn.metrics	- Confusion_matrix was imported to tabulate the confusion matrix comparing the number of correctly and wrongly predicted values. - This was used to visualize results and interpret how the data points were classified with the created model.
Accuracy_score from sklearn.metrics	- Accuracy_score was imported to calculate the accuracy of the model created by comparing the predicted output to the actual output.

Table 1: List of Python Library Package installed for implementation.

4. SUITABLE DATASET SOURCING AND MANAGEMENT

Based on the inputs and output parameters selected in Part 1 of this study, a suitable dataset was sourced from UCI Machine Learning Repository from the dataset titled "Heart Disease Dataset" [1]. The dataset was sourced from the Cleveland Clinic Foundation, Hungarian Institute of Cardiology, V.A. Medical Center and University Hospital of Zurich. The dataset consists 13 columns of different features / input parameters and a final column of the output. The nature of each input parameter relevant to this study and the output is detailed in Table 2 below.

Data-frame feature column	Explanation
Age	- Numerical values representing age in years.
Chest Pain (cp)	- Categorical form of data represented in 4 different numerical values. - Each number represents the type of chest pain.
Resting Systolic Blood Pressure (trestbps)	- Numerical values representing the systolic blood pressure of the individual in mmHg.
Total Serum Cholesterol (chol)	- Numerical values representing the total serum cholesterol of the individual in mg/dl.
Fasting Blood Sugar (fbs)	- Categorical form of data represented in binary values of 0 or 1. - 1 represents blood sugar level >120 mg/dl and 0 represents blood sugar level ≤ 120 mg/dl.
Output (num)	- Categorical form of data represented in 5 different numerical values of 0, 1, 2, 3 and 4. - An increase from 1 category to the other indicates a higher-level risk for an individual to be susceptible to heart disease. - 0 – Healthy, 1 – Mild/Sick 1, 2 – Moderate/Sick 2, 3 – Severe/Sick 3 and 4 – Very Severe/Sick 4.

Table 2: Explanation of Data Frame inputs and output

The raw dataset was first imported into the program from the project repository using the pd.read_csv() function. 3 different .csv files were obtained from [1] consisting of a total of 720 input data points and their respective level of heart disease risk. The pd.concat() function was used to concatenate all these 3 groups of datasets together into a single data frame. The next step involved filtering inputs features / columns within the dataset which were relevant to the study.

From analysis of the raw data, it was noticed that there were numerous missing points from the raw dataset. Missing data were filled in with "?" character. Hence, each row of the raw dataset was iterated through and rows that consisted of "?" value was removed from the dataset using the df.drop() function. This was executed to ensure that only rows with all input features value filled in were considered for the study. Out of the 720 datapoints, 108 rows of data were removed leaving a remaining of 612 rows of data. These remaining 612 data points were then used for evaluating / testing the membership functions, rules, and the entire fuzzy logic system model, as detailed further in this report.

5. MODELLING INPUTS AND OUTPUT USING MEMBERSHIP FUNCTIONS

This section details the steps taken to define the membership functions for each linguistic representation class of the 5 inputs and the output considered in the diagnosis of heart disease that were discussed in Part 1 of this study.

5.1 Defining the Membership Functions

Membership functions are defined as mathematical functions that specifies the degree to which a given input belongs to a certain set and has a value always limited between 0 and 1. To accurately define the membership functions for the inputs and output, the characteristics of each input parameter, their range of values and their fuzzy linguistic representation were analyzed. In addition, membership functions implemented in literatures [2][3][4][5][6] as discussed in Part 1 were also taken as reference for implementing the membership functions.

Depending on the nature of the of input parameter, one (1) or more sets of membership functions were assigned to the respective inputs to calculate its corresponding fuzzy membership. For example, the three classes of low + medium + high for cholesterol were represented using R-function (Decreasing function) + Triangular function + L-Function (Increasing Function) respectively as one set of membership functions and Decreasing Sigmoid + Gaussian + Increasing Sigmoid membership functions respectively as another. These different sets of membership functions were only implemented for inputs with numerical data. This was done to see how the variations of membership functions affected the accuracy of the model and to choose the membership function that best suited the input. For inputs that were categorical in nature such as chest pain and fasting blood sugar, only a single set of membership function that best represented the data was defined with conditions to exceptions implemented to convert any fuzzy value into a crisp categorical data. Further elaboration will be detailed under each individual input sections.

These different combinations of membership functions were plotted. It was ensured that the different sets of membership functions created were comparable to one another and not significantly different from one another. In addition, the membership functions also respected the defined fuzzy linguistic representations.

Similar steps were also taken when implementing the memberships for the output. It is important to note that the membership function and fuzzy linguistic representation of the output variable was first defined before proceeding with the inputs. This is to ensure that the there is a consistent output representation across all input for a standardized evaluation.

5.2 Evaluation of Membership Function and Final Membership Selection

It is important to note that the Mamdani model was selected for implementation of this study. The justification for Mamdani model selection is detailed in Section 7.1.

The fuzzy membership for each class of the dataset points were then computed using the different sets of membership functions defined. Subsequently, several fuzzy rules were established for each input to test the membership functions' performance. The rules are detailed in Section 6.1. These rules will also be utilized in the finalization of rules for the entire model.

Once the rules were established, an aggregation (union) of the rules were calculated using the np.maximum() function. This would then produce the fuzzy output value. Defuzzification was then carried out by calculating the centroid of the area under the fuzzy set with respect to the output. The formula for defuzzification is shown below in Equation 1.

$$\overline{\boldsymbol{x}} = \frac{\int \boldsymbol{\mu}(\boldsymbol{x})_{\boldsymbol{output}} * \boldsymbol{x}\, \boldsymbol{dx}}{\int \boldsymbol{\mu}(\boldsymbol{x})_{\boldsymbol{output}}\, \boldsymbol{dx}}$$

Equation 1: Equation to calculate centroid of area under output membership function curve

Where $\overline{\boldsymbol{x}}$ is the defuzzied output, $\boldsymbol{\mu}(\boldsymbol{x})_{\boldsymbol{output}}$ is the membership function of the output and $\boldsymbol{x}$ is the aggregated rules matrix.

The output values from the dataset are defined as categorical values of 0, 1, 2, 3 and 4. The defuzzied output from the Mamdani model however produces an output of numerical decimal values between 0 – 4 (min and max range of the membership function defined for the output of this study). Considering this and the explanation provided in Section 5.8 on the output, the predicted and dataset output were regrouped as per Table 13.

Once the predicted and actual output were categorized to their respective final classes, the confusion matrix and the accuracy of the Mamdani Model for the inputs was then evaluated. The confusion matrix and the accuracy were observed parameters used to gauge and evaluate the performance of the model with different membership functions' parameters and rules for each input. Ultimately, the Mamdani Model with the membership function combination that produced the highest accuracy was selected as the final membership function set to represent that particular input.

5.3 Chest Pain

As discussed in Part 1 of the assignment, chest pain is categorized into 4 different fuzzy linguistic categories of 1 – Typical Angina, 2 – Atypical Angina, 3 – Non-Typical Angina, and 4 – Asymptomatic.

As the input for chest pain from the dataset is categorical in nature and not continuous, only a single set of membership function was implemented. This set of membership function comprised of 4 triangle functions representing each category of chest pain. The functions have a peak value of 1, 2, 3 and 4 the function increases and decreases linearly at a distance of ± 0.5 from the peak. A visualization of the membership function is depicted in Figure 2.

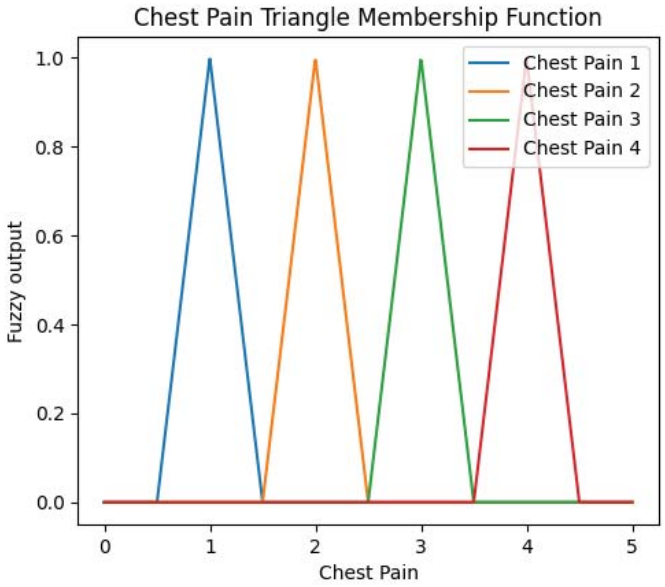


Figure 2: Visualization of membership function for chest pain

Additionally, once the fuzzy membership values were calculated, a condition was set in the program to categorize each element in the fuzzy sets to their respective categories. This is done by rounding off any values in the fuzzy membership to 1 if only its value is greater than 0. Table 3 depicts the categorization method.

Categorized input.	Rule for categorization
Triangle Function - Typical Angina (1)	IF μ (x) > 0 THEN μ (x) = 1
Triangle Function - Atypical Angina (2)	
Triangle Function - Non-Anginal Pain (3)	
Triangle Function - Asymptomatic (4)	

Table 3: Condition set for redefining the membership of age into square function

By implementing the above step, the triangle membership functions were then converted into square membership functions as depicted in Figure 3.

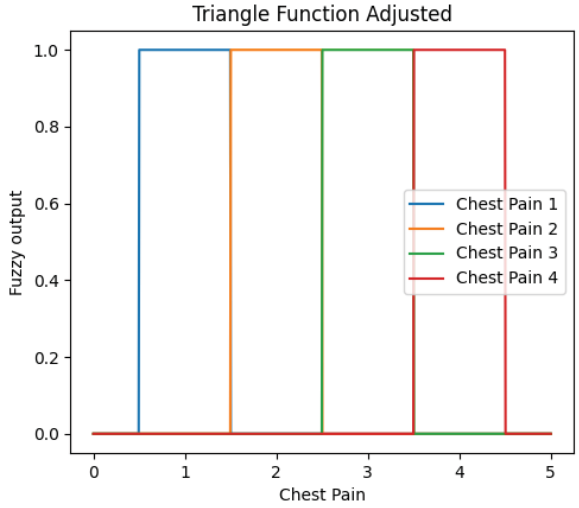


Figure 3: Visualization of membership function for chest pain after applying conditions in Table 4

5.4 Cholesterol

As discussed in Part 1 of the assignment, Cholesterol is categorized into 3 different fuzzy linguistic categories of Low (Cholesterol level < 200 mg/dl), Medium (194-245 mg/dl) and High (>240 mg/dl).

The input data for Cholesterol is numerical. Hence, 3 different sets of membership functions were created to represent the data. The 3 sets of membership functions for respective classes of Cholesterol level are defined in Table 4 below:

	Linguistic Class	**Membership Function**	**Equation**
Set 1	Low	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 151 \\ \frac{200 - x}{200 - 151}, 151 \leq x \leq 200 \\ 0, x > 200 \end{cases}$
	Medium	Triangle Function	$\mu(x) = \begin{cases} 0, x \leq 194 \\ \frac{x - 194}{219 - 194}, 194 < x \leq 219 \\ \frac{245 - x}{245 - 219}, 219 < x < 245 \\ 0, x \geq 245 \end{cases}$
	High	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 240 \\ \frac{x - 240}{263 - 240}, 240 \leq x \leq 263 \\ 1, x > 263 \end{cases}$
Set 2	Low	Decreasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{0.5(x-176)}}, a = -0.5$
	Medium	Gaussian	$\mu(x) = e^{-\frac{(x-219)^2}{2*10^2}}$
	High	Increasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{-0.5(x-252)}}, a = 0.5$
Set 3	Low	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 151 \\ \frac{200 - x}{200 - 151}, 151 \leq x \leq 200 \\ 0, x > 200 \end{cases}$
	Medium	Gaussian Function	$\mu(x) = e^{-\frac{(x-219)^2}{2*10^2}}$
	High	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 240 \\ \frac{x - 240}{263 - 240}, 240 \leq x \leq 263 \\ 1, x > 263 \end{cases}$

Table 4: Table Summarizing the proposed sets of membership functions for Cholesterol

A visualization of the 3 sets of membership function is depicted in Figure 4.

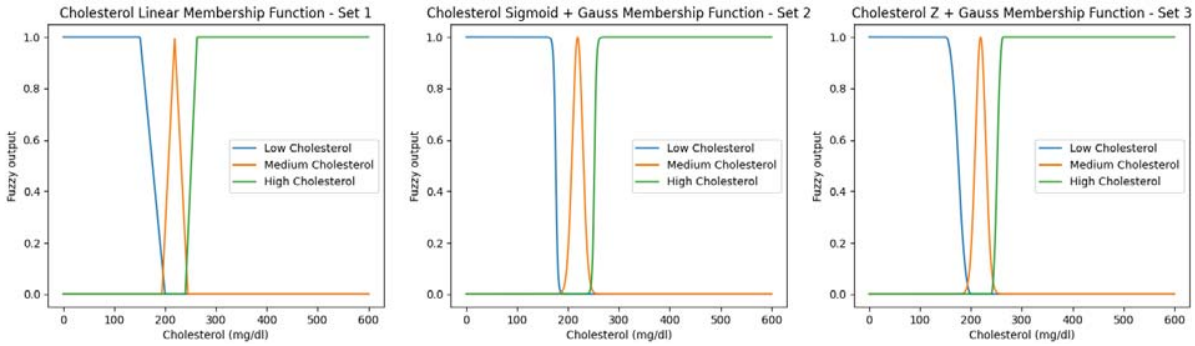


Figure 4: Visualization of 3 sets of membership function for Cholesterol:

Figure 5 Shows that the 3 sets of membership functions defined for testing do not variate significantly from one another and are comparable:

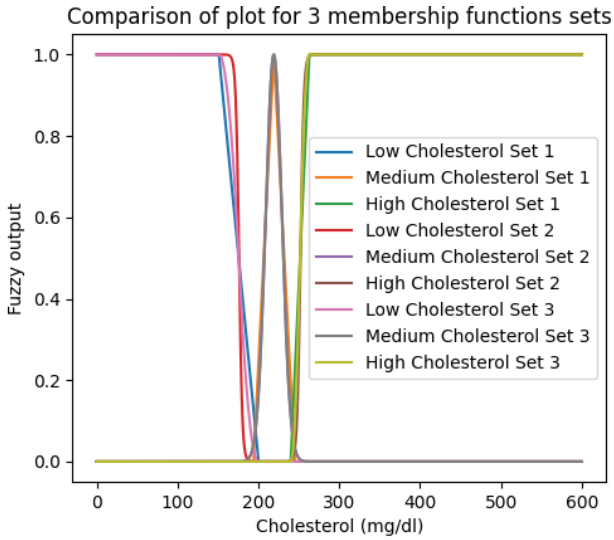


Figure 5: Visualization of all 3 sets of membership functions plotted together for Cholesterol

The input values from the dataset were evaluated using the membership functions defined in Table 4**.** 3 rules were the created and evaluated. The rules are:

- If Cholesterol Level is Low, then the individual is either Healthy or Sick 1
- If Cholesterol Level is Medium, then the individual is either Sick 1 or Sick 2.
- If Cholesterol Level is High, then the individual is either Sick 3 or Sick 4

A union aggregation of these rules was calculated and the defuzzied output was calculated. The output was then categorized to its respective output categories and the confusion matrix and accuracy was calculated. Based on the results obtained, Set 1 predicted the output with an accuracy of 51.31%, Set 2 predicted the output with an accuracy of 52.61% and Set 3 predicted the output with an accuracy of 51.80%. Evaluating based on accuracy, Set 2 was selected as the final set of membership functions to represent the input of Cholesterol level.

5.5 Blood Pressure

As discussed in Part 1 of the assignment, Blood Pressure is categorized into 3 different fuzzy linguistic categories of Low (Blood Pressure < 125 mm/Hg), Medium (120-140 mm/Hg) and High (>130 mm/Hg).

The input data for Blood Pressure is numerical. Hence, 3 different sets of membership functions were created to represent the data. The 3 sets of membership functions for respective classes of Blood Pressure level are defined in Table 5 below.

	Linguistic Class	**Membership Function**	**Equation**
Set 1	Low	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 111 \\ \frac{125 - x}{125 - 111}, 111 \leq x \leq 125 \\ 0, x > 125 \end{cases}$
	Medium	Triangle Function	$\mu(x) = \begin{cases} 0, x \leq 120 \\ \frac{x - 120}{130 - 120}, 120 < x \leq 130 \\ \frac{140 - x}{140 - 130}, 130 < x < 140 \\ 0, x \geq 140 \end{cases}$
	High	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 130 \\ \frac{x - 130}{157 - 130}, 130 \leq x \leq 157 \\ 1, x > 157 \end{cases}$
Set 2	Low	Decreasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{0.5(x-118)}}, a = -0.5$
	Medium	Gaussian	$\mu(x) = e^{-\frac{(x-130)^2}{2*5^2}}$
	High	Increasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{-0.5(x-144)}}, a = 0.5$
Set 3	Low	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 111 \\ \frac{125 - x}{125 - 111}, 111 \leq x \leq 125 \\ 0, x > 125 \end{cases}$
	Medium	Gaussian Function	$\mu(x) = e^{-\frac{(x-130)^2}{2*5^2}}$
	High	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 130 \\ \frac{x - 130}{157 - 130}, 130 \leq x \leq 157 \\ 1, x > 157 \end{cases}$

Table 5: Table Summarizing the proposed sets of membership function for Blood Pressure

A visualization of the 3 sets of membership function is depicted in Figure 6.

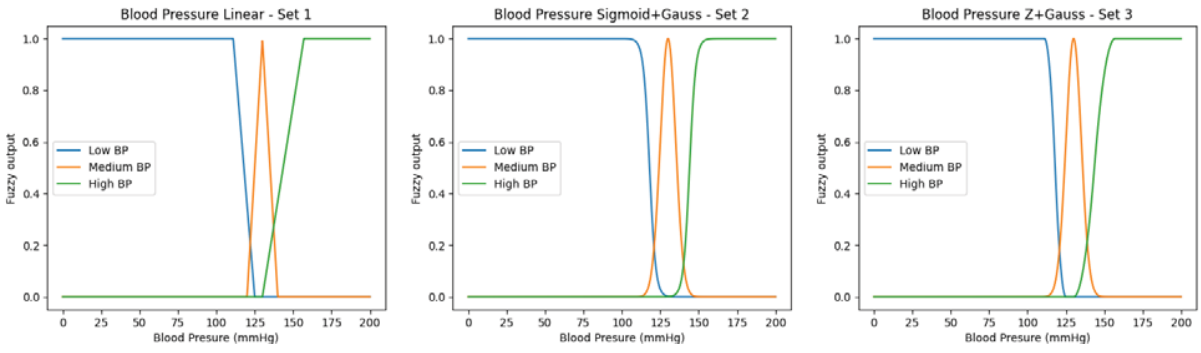


Figure 6: Visualization of 3 sets of membership function for Blood Pressure

Figure 7 Shows that the 3 sets of membership functions defined for testing do not variate significantly from one another and are comparable.

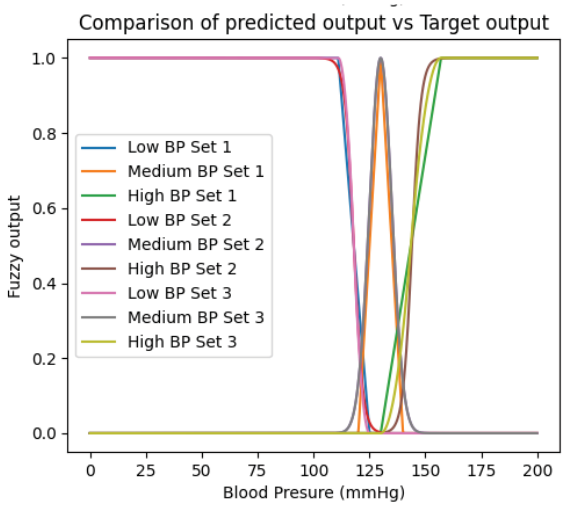


Figure 7: Visualization of all 3 sets of membership functions plotted together for Blood Pressure

The input values from the dataset were evaluated using the membership functions defined in Table 5**.** 3 rules were the created and evaluated. The rules are:

- If Blood Pressure Level is Low, then the individual is either Healthy or Sick 1.
- If Blood Pressure Level is Medium, then the individual is either Sick 1 or Sick 2.
- If Blood Pressure Level is High, then the individual is either Sick 3 or Sick 4.

A union aggregation of these rules was calculated and the defuzzied output was calculated. The output was then categorized to its respective output categories and the confusion matrix and accuracy was calculated. Based on the results obtained, Set 1 predicted the output with an accuracy of 61.11%, Set 2 predicted the output with an accuracy of 72.71% and Set 3 predicted the output with an accuracy of 61.11%. Set 2 was selected as the final set of membership functions to represent the input of Blood Pressure level.

5.6 Blood Sugar

As discussed in Part 1 of the assignment, Blood Sugar is categorized into 2 different fuzzy linguistic categories of Low (Blood Sugar ≤ 120 mm/Hg), and High (Blood Sugar >130 mm/Hg).

The input data for Blood Sugar is categorical in nature and is presented in the form of binary classes of 0 and 1. While the input data defers from the identified fuzzy linguistic categories in part 1 of the study, the fuzzy linguistic categories can be converted into Low Blood Sugar = 0 and High Blood Sugar = 1. This conversion segregates the Blood Sugar as either high, or not high (low). Due to its binary nature, no alternative membership function was defined for testing. The membership function for Blood Sugar is depicted in Figure 9.

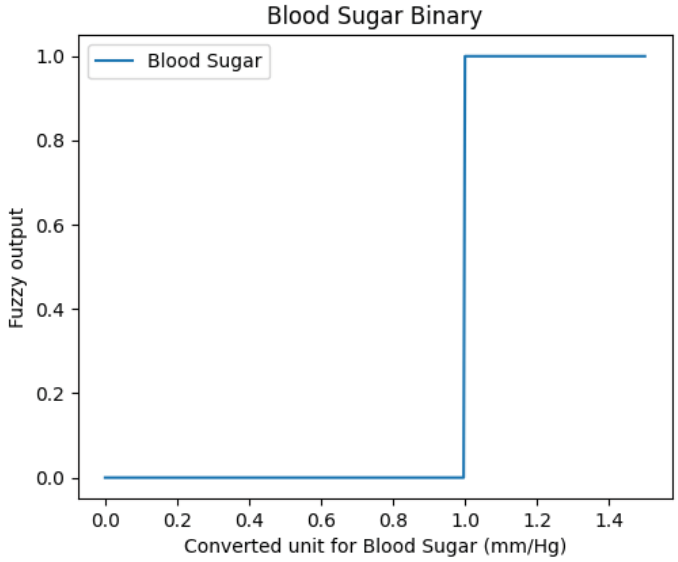


Figure 8: Visualization of membership function for Blood Sugar with single membership

Initially, a single membership function was only adopted to represent the fuzzy linguistic categories, whereby the data point <0 implies $\mu_{sugar} = 0$, and data point ≥ 1 is represented by $\mu_{sugar} = 1$ (High Blood Sugar). Although the graph above, Figure 8 logically resembles the fuzzy linguistic categories, for cases where the individual is low in blood sugar level, the membership of that individual will be 0. This consequently leads to Not A Number ("**NAN**") errors at later stages in defuzzification when calculating the area under the centroid as membership values are 0. As an alternative to the exception to resolve the error and as an additional flexibility to define rules, an additional membership function for low blood sugar was defined making 2 membership functions for blood sugar level of high and low blood sugar. The updated membership is depicted in Figure 9.

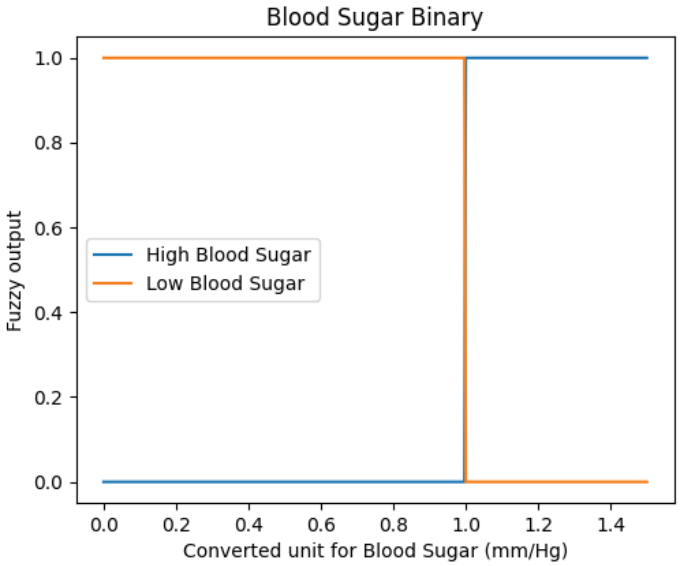


Figure 9: Visualization of membership function for Blood Sugar with dual membership

The categories were segregated into 2 combined binary categories, where low Blood Sugar returns 1.0 for its membership value when the converted unit for Blood Sugar is <1.0 but returns 0 for its membership value when the converted unit for Blood Sugar is ≥ 1.0. On the other hand, high Blood Sugar shares inverse relationship with low Blood Sugar, where the membership function returns 1.0 when converted unit for Blood Sugar is ≥ 1.0 and the membership function returns 0 when converted unit for Blood Sugar is <1.0. With the above modification, the **"NAN"** error is therefore resolved, as each data point will return a membership value which is greater than 0.

5.7 Age

As discussed in Part 1 of the assignment, Age is categorized into 3 different fuzzy linguistic categories of Young (Age < 33 years old), Middle (33-45 years old) and Old (>45 years old).

The input data for Age is numerical. Therefore, 3 different sets of membership functions were created to represent the data. The 3 sets of membership functions for respective classes of Age are defined in Table 6 below. The 3rd set of membership function adopted linear method. Similar to 1st set of membership function, the only difference is the point of separation / overlapping of age between classes for each membership function, which will be discussed further below.

	Linguistic Class	**Membership Function**	**Equation**
Set 1	Young	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 27 \\ \frac{33 - x}{33 - 27}, 27 \leq x \leq 33 \\ 0, x > 33 \end{cases}$
	Mid	Triangle Function	$\mu(x) = \begin{cases} 0, x \leq 32 \\ \frac{x - 32}{39 - 32}, 32 < x \leq 39 \\ \frac{46 - x}{46 - 39}, 39 < x < 46 \\ 0, x \geq 46 \end{cases}$
	Old	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 45 \\ \frac{x - 45}{51 - 45}, 45 \leq x \leq 51 \\ 1, x > 51 \end{cases}$
Set 2	Young	Decreasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{3(x-30)}}, a = -3$
	Mid	Gaussian	$\mu(x) = e^{-\frac{(x-39)^2}{2*2^2}}$
	Old	Increasing Sigmoid	$\mu(x) = \frac{1}{1 + e^{-3(x-48)}}, a = 3$
Set 3	Young	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < 27 \\ \frac{33 - x}{33 - 27}, 27 \leq x \leq 33 \\ 0, x > 33 \end{cases}$
	Mid	Triangle Function	$\mu(x) = \begin{cases} 0, x \leq 33 \\ \frac{x - 33}{39 - 33}, 33 < x \leq 39 \\ \frac{44 - x}{44 - 39}, 39 < x < 44 \\ 0, x \geq 44 \end{cases}$
	Old	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < 240 \\ \frac{x - 240}{263 - 240}, 240 \leq x \leq 263 \\ 1, x > 263 \end{cases}$

Table 6:Table Summarizing the proposed sets of membership function for Age

A visualization of the 3 sets of membership function is depicted in Figure 10:

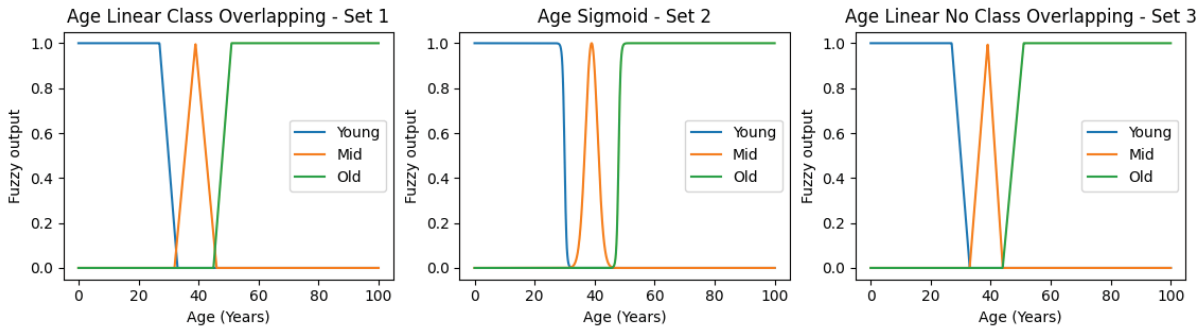


Figure 10: Visualization of 3 sets of membership function for Age

The difference between 1st set and 3rd set of membership function is the overlapping between classes in Set 1 and no overlapping of classes in Set 3. The 3rd set of membership functions shows a clear segregation of category between each age class, which is desired as it logically reflects the actual Age categorization. However, by implementing Set 3, there are points at age = 33 and age = 45 where there is no-memberships for these 2 age values at the break-off point between young-mid and mid-old.

When the membership for the input is zero, an error occurs indicating that these 2 ages do not belong to the universal set which is not accurate. Hence to mitigate this, a condition to the exception was created whereby when age = 33, $\mu(33) = 0.5*\min(\mu(x)_{young})+ 0.5*\min(\mu(x)_{mid})$ and when age = 45, $\mu(45) = 0.5*\min(\mu(x)_{mid})+ 0.5*\min(\mu(x)_{old})$ based on the input dataset.

Based on the 1st set of membership function, a datapoint with an Age of 32 years old is considered as Young and Middle, which is not clear and undesired. It is important to note that the 3 membership functions defined as shown in Figure 11 do not variate significantly from one another and are comparable.

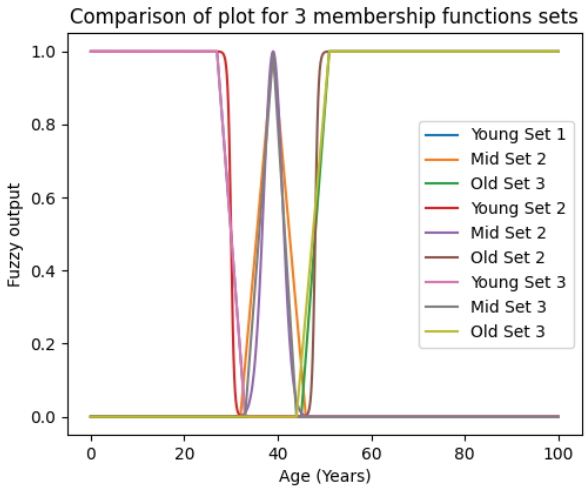


Figure 11: Visualization of all 3 sets of membership functions plotted together for Age

The input values from the dataset were evaluated using the membership functions defined in Table 6. 3 rules were the created and evaluated. The rules are:

- If Age is Low, then the individual is either Healthy or Sick 1
- If Age is Medium, then the individual is either Sick 1 or Sick 2.
- If Age is High, then the individual is either Sick 3 or Sick 4

A union aggregation of these rules was calculated and the defuzzied output was calculated. The output was then categorized to its respective output categories and the confusion matrix and accuracy was calculated. Based on the results obtained, all 3 sets predicted the output with an accuracy of 88.89%. Since all 3 sets shares the same accuracy, the basis of selection was narrowed down to the logical separation when identifying the membership function. Therefore, Set 3 was selected as the final set of membership functions to represent the input of Age given the reasons elaborated above.

5.8 Output

As discussed in Part 1 of the assignment, the Output is categorized into 5 different fuzzy linguistic categories of Healthy, Sick 1, Sick 2, Sick 3 and Sick 4. The definition of the output membership occurs in 2 stages, before regrouping and after regrouping.

As the output from the dataset is categorical in nature and not continuous, only a single set of membership function was implemented. This set of membership function comprised of decreasing function for healthy starting with membership value of 1 which then reduces linearly to 0 starting from x = 0.25 - 0.5. An increasing function is used for Sick 4 membership starting from a membership of 0 from x = 0 - 3.5 which increases linearly to 1 from x = 3.5-3.75. The remaining 3 categories used triangle functions representing each category of the output. The functions have a peak value of 1, 2, and 3 and the function increases and decreases linearly at a distance of ± 0.5 from the peak. A visualization of the membership function is depicted in Figure 12.

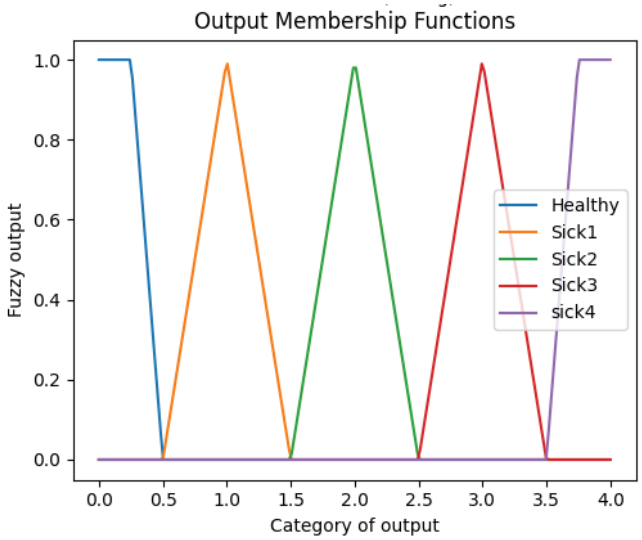


Figure 12: Visualization initial membership function for Output

The equation that mathematically represent the output is shown in Table 7

Linguistic Class	Membership Function	Equation
Healthy	R-Function (Decreasing Function)	$\mu(x) = \begin{cases} 1, x < c \\ \frac{d-x}{d-c}, c \leq x \leq d \\ 0, x > d \end{cases}$
Sick 1, Sick 2 and Sick 3	Triangle Function	$\mu(x) = \begin{cases} 0, x \leq a \\ \frac{x-a}{m-a}, a < x \leq m \\ \frac{b-x}{b-m}, m < x < b \\ 0, x \geq b \end{cases}$ Where m = 1, 2 or 3 a = m-0.5 & b = m+0.5
High	L-Function (Increasing Function)	$\mu(x) = \begin{cases} 0, x < a \\ \frac{x-a}{b-a}, a \leq x \leq b \\ 1, x > c \end{cases}$

Table 7: Table Summarizing initial membership function for Output

The defuzzied outputs from the model in the form of numerical decimal values were initially categorized into their respective categories based on the conditions set in Table 13.

Output	Range of Values
Healthy	<0.5
Sick 1	≥ 0.5 and <1.5
Sick 2	≥ 1.5 and <2.5
Sick 3	≥ 2.5.5 and 3.5
Sick 4	≥ 3.5

Table 8: Initial Conditions for Output Categorization (Before regrouping)

By implementing the above step, the triangle membership functions were then converted into square membership functions as depicted in Figure 13.

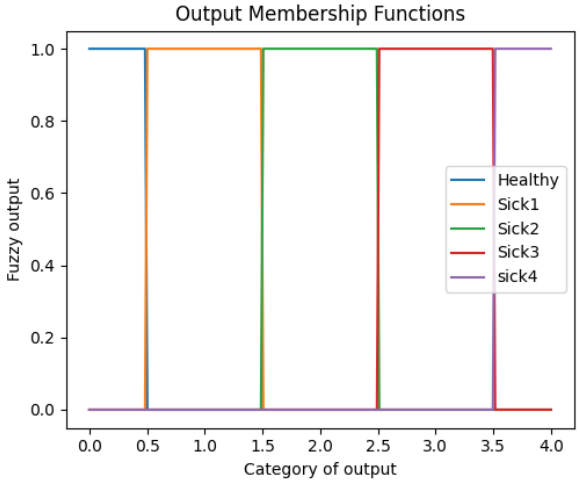


Figure 13: Visualization of membership function of output after setting conditions to the triangle function (before regrouping the output categories)

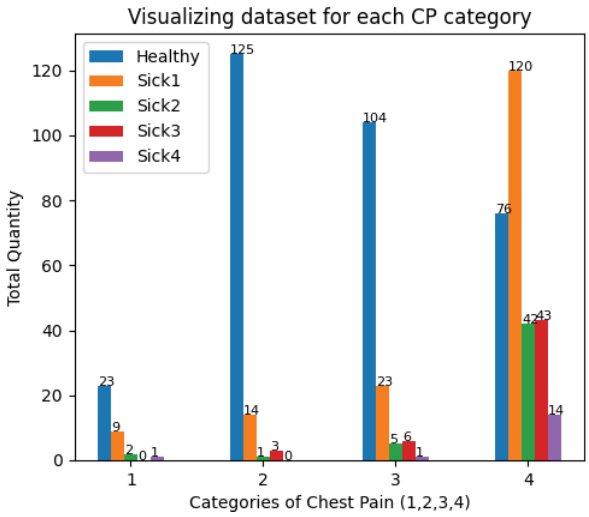
This was the initial standardized output used to evaluate the memberships of the inputs.

However, based on initial stages of testing on individual inputs, it was found that the accuracy of the model was very low with an accuracy of only within 10 - 30%. Several approaches were taken to improve the accuracy of the outputs which included adjusting the membership function parameters / range as well as trying different combinations of rules. However, the accuracy of the results still did not significantly improve. The initial accuracies obtained for the inputs are shown below in Table 9.

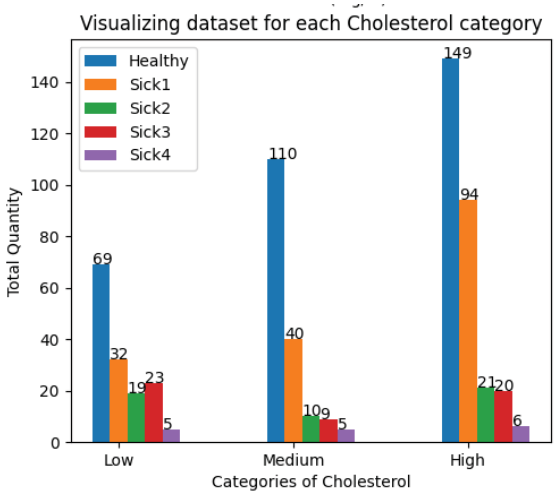
Membership Set	**Accuracy (%) - No regrouping of output**
Chest Pain	
Set 1	23.86
Cholesterol	
Set 1	10.13
Set 2	9.31
Set 3	9.97
Blood Pressure	
Set 1	14.22
Set 2	16.50
Set 3	14.22
Blood Sugar Level	
Set 1	50.49
Age	
Set 1	28.10
Set 2	27.78
Set 3	28.10

Table 9: Initial accuracies of model for individual inputs before regrouping (5 tier output categories)

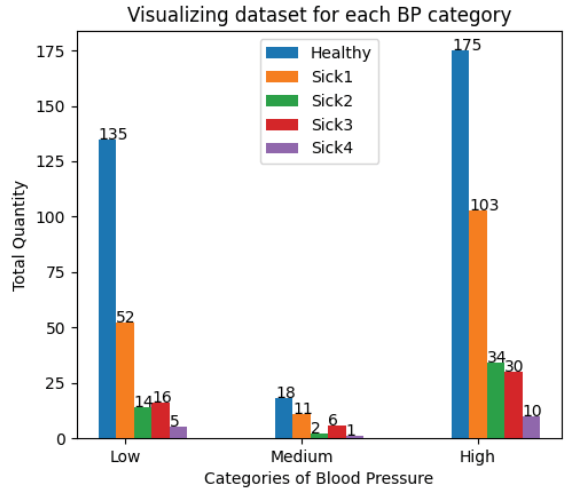
Hence, the alternative to make adjustments to the output categories was considered. Prior to making any adjustments, the distribution of output values for each fuzzy class of input was visualized. This was done by calculating the total number of datapoints for each output classes for all inputs and segregating them according to each fuzzy input class. The is visually shown in Figure 14 (a) to Figure 14 (e). From the bar plots, it can be observed that a majority of the output for the datapoints in the dataset for almost all of the input classes are skewed towards the healthy side (left side). In essence, there is an imbalance distribution of output data where there are a lot of datapoints with individuals that are healthy and less datapoints of individuals who are sick particularly on levels Sick 3 and Sick 4. This means that the dataset is more biased towards the healthy class as opposed to the rest due to the high purity of "Healthy" class in the data.



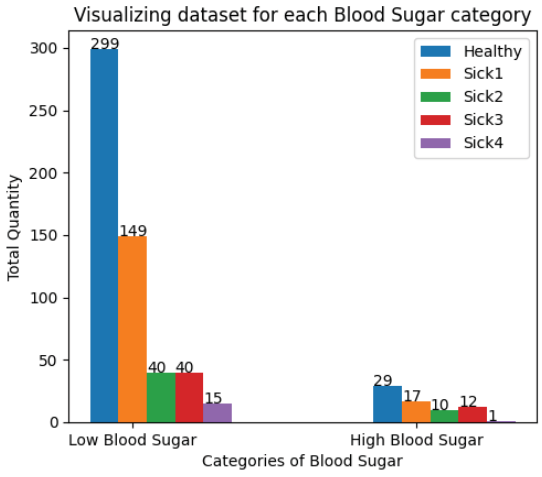
(a) Distribution of output for each level of chest pain



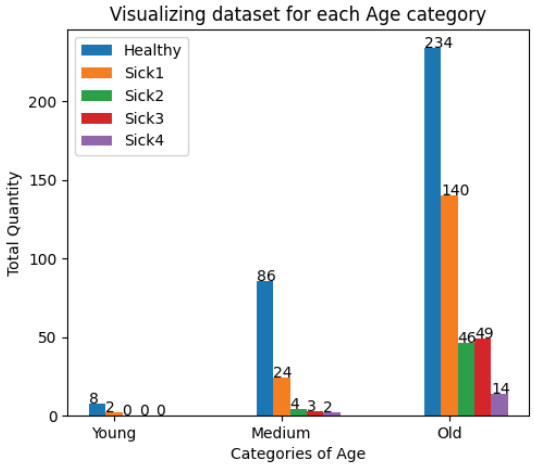
(b) Distribution of output for each level of Cholesterol



(c) Distribution of output for each level of Blood Pressure



(d) Distribution of output for each level of Blood Sugar



(e) Distribution of output for each level of Age

Figure 14:Distribution of output for each input parameter

Another observation made was that there were too many output categories. From the Confusion Matrix plotted for all of the inputs using the initial 5 categories of outputs, it was observed that a lot of Healthy output were misclassified as Sick 1 or Sick 2 or Sick 1 were misclassified as Sick 2 and vice versa. An example can be seen for Cholesterol and Blood Pressure using Set 2 that were selected for both these inputs in Table 10 and Table 11.

Predicted Output / Actual output	0 – Healthy	1 – Sick 1	2 – Sick 2	3 – Sick 3	4 – Sick 4
0 – Healthy	0	57	140	131	0
1 – Sick 1	0	28	48	90	0
2 – Sick 2	0	20	9	21	0
3 – Sick 3	0	24	8	20	0
4 – Sick 4	0	5	6	5	0

Table 10: Confusion Matrix for Cholesterol

Predicted Output / Actual output	0 – Healthy	1 – Sick 1	2 – Sick 2	3 – Sick 3	4 – Sick 4
0 – Healthy	0	153	119	56	0
1 – Sick 1	0	63	56	47	0
2 – Sick 2	0	16	21	13	0
3 – Sick 3	0	22	13	17	0
4 – Sick 4	0	6	3	7	0

Table 11: Confusion Matrix for Blood Pressure

In order to improve accuracy, the 5 different categories of output were further simplified / regrouped into 3 categories as detailed below in Table 12. The Fuzzy range for grouping is defined as Table 13 below. The visualization of the regrouped output membership is depicted in Figure 15.

Initial Output Category	Final Output Category
0 - Healthy	0 – Low Risk of Heart Disease *Justification: Healthy, Sick 1 and Sick 2 are low level risk groups. Sick 2 is borderline risk level to medium.
1 – Sick 1	
2 – Sick 2	
3 – Sick 3	1 – Medium Risk of Heart Disease *Justification: Sick 3 can be defined as medium to borderline high
4 – Sick 4	2 – High risk of Heart Disease *Justification: Sick 4 represents the highest level of risk.

Table 12: Conditions to regroup 5 tier output into 3 tier output

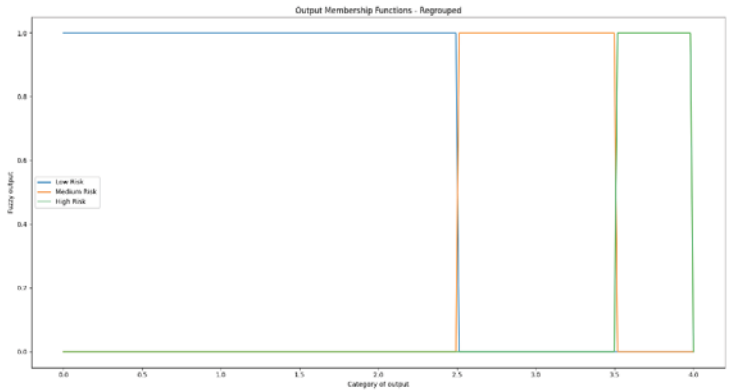


Figure 15: Visualization of the output membership functions after regrouping (3 tier)

Output	Range of Values
Low Heart Disease Risk	<2.5
Medium Heart Disease Risk	≥ 2.5 and <3.5
High Heart Disease Risk	≥ 3.5

Table 13: Fuzzy range definition for regrouped output with 3 tiers

By regrouping the output categories, the accuracy of predicted outputs by the model significantly improved as shown in Table 14.

Membership Set	Accuracy (%) - No regrouping of output	Accuracy (%) - With regrouping of output
Chest Pain		
Set 1	23.86	61
Cholesterol		
Set 1	10.13	51.31
Set 2	9.31	52.61
Set 3	9.97	51.80
Blood Pressure		
Set 1	14.22	61.11
Set 2	16.50	72.71
Set 3	14.22	61.11
Blood Sugar Level		
Set 1	50.49	88.89
Age		
Set 1	28.10	88.89
Set 2	27.78	88.89
Set 3	28.10	88.89

Table 14: Accuracies of model for individual inputs after regrouping (3 tier output categories)

By regrouping the outputs, a clearer distinction can be made between each output class. It is important to note that this step is only done at the final stage of the program after the defuzzied output values were calculated by the model and when plotting the confusion matrix and accuracy calculation and is applied for all inputs. This is done using a self-built function called roundup().

A similar method of classifying the output was also implemented by [5][6][7][8][9] as well where the outputs were defined in 2 or 3 categories only.

5.9 Summary of Membership Functions

A summary of the membership functions defined for each input and output parameter is simplified in Table 15 below.

Parameter	Membership Functions Selected
Chest Pain	- Triangular Function for all fuzzy class - Condition to categorize Fuzzy output to respective category
Cholesterol	- Low – Sigmoid Decreasing - Medium – Gaussian - High – Sigmoid Increasing
Blood Pressure	- Low – Sigmoid Decreasing - Medium – Gaussian - High – Sigmoid Increasing
Blood Sugar level	- Low - A spike function of dropping from 1 to 0 at 120 mg/dl - High - A spike function of increasing from 1 to 0 at 120 mg/dl
Age	- Young – R Function (Decreasing) - Middle – Triangular Function - Old – L Function (Increasing)
Output	- Healthy – R Function (Decreasing) - Sick 1, Sick 2 and Sick 3 – Triangle - Sick 4 – L Function (Increasing) - 5 categories were regrouped into 3 categories of Low Risk (Healthy, Sick 1 and Sick 2), Medium Risk (Sick 3) and High Risk (Sick 4)

Table 15: Summary of membership functions for inputs and output

6. DEFINITION OF RULES

Once membership functions were defined, fuzzy rules were then established for the model. A total of 11 rules were defined and tested.

At prior stages where testing of potential membership functions was carried out, simple rules were created specific for each input as mentioned in Section 5.2. These rules were then carried forward to and used as the building blocks for the 11 rules of the entire fuzzy system discussed in this section. The 11 rules consist of different combinations of the simple rules. Several factors were taken into consideration when creating the 11 fuzzy rules. These rules were developed by means of reference from literatures and by visualizing the distribution of output data for each class of input (i.e., low, medium, and high). The rules were represented linguistically by means of implication statements (IF (PRECEDENT) THEN (CONSEQUENT) statements) and mathematically calculated by means of compositions of the membership function matrix for the particular input class and the output matrix (using np.fmin() function).

These 11 rules were created and tested individually before integrating 9 out of the 11 to finalize the rules for the finalized model.

6.1 Building Simple Rules for each Input

The simple rules built for each input parameter during testing stage is defined in Table 16 as below:

Input - Chest Pain		**Accuracy (%)**
Rule 1	IF (chest pain = 4), THEN (Heart Disease Risk = Healthy or Sick 1)	61
Rule 2	IF (chest pain = 3), THEN (Heart Disease Risk = Healthy or Sick 1 or Sick 2)	
Rule 3	IF (chest pain = 2), THEN (Heart Disease Risk = Sick 3 or Sick 4)	
Rule 4	IF (chest pain = 1), THEN (Heart Disease Risk = Sick 3 or Sick 4)	
Input - Cholesterol		**Accuracy (%)**
Rule 5	IF (cholesterol level = Low), THEN (Heart Disease Risk = Healthy or Sick 1)	52.61
Rule 6	IF (cholesterol level = Medium), THEN (Heart Disease Risk = Sick 1 or Sick 2)	
Rule 7	IF (cholesterol level = High), THEN (Heart Disease Risk = Sick 3 or Sick 4)	
Input - Blood Pressure		**Accuracy (%)**
Rule 8	IF (Blood Pressure = Low), THEN (Heart Disease Risk = Healthy or Sick 1)	72.71
Rule 9	IF (Blood Pressure = Medium), THEN (Heart Disease Risk = Sick 1 or Sick 2)	
Rule 10	IF (Blood Pressure = High), THEN (Heart Disease Risk = Sick 3 or Sick 4)	

Input - Blood Sugar		**Accuracy (%)**
Rule 11	IF (Blood Sugar= Low), THEN (Heart Disease Risk = Healthy)	88.89
Rule 12	IF (Blood Sugar = High), THEN (Heart Disease Risk = Sick 1 or Sick 2 or Sick 3 or Sick 4)	
Input - Age		**Accuracy (%)**
Rule 13	IF (Age = Young), THEN (Heart Disease Risk = Healthy)	88.89
Rule 14	IF (Age = Middle), THEN (Heart Disease Risk = Sick 1 or Sick 2)	
Rule 15	IF (Age = Old), THEN (Heart Disease Risk = Sick 3 or Sick 4)	

Table 16: Rules for each input

6.2 Definition of rules for the final Mamdani Model:

11 rules were defined as shown in Table 17 to be implemented in the final Mamdani Model. The 9 highest producing rules are selected for the final rule aggregation.

No	**Chest pain (cp1, cp2, cp3,cp4)**	**Cholesterol (LOW, MEDIUM, HIGH)**	**Blood Pressure (LOW, MEDIUM, HIGH)**	**Blood Sugar (LOW, HIGH)**	**Age (YOUNG, MIDDLE, OLD)**	**Output (LOW, MEDIUM, HIGH)**	**Accuracy (%)**
1	1	high	high	N/A	middle OR old	medium OR high	83.50
IF (Middle OR Old age) AND (Chest pain 1) AND (BP High OR Chol High) THEN (Medium OR High Risk)							
2	4	low OR medium	low OR medium	N/A	middle OR young	Low	88.89
IF (Middle OR Young) AND (chest pain 4) AND ((BP low OR BP medium) OR (Chol low OR Chol medium)) THEN (Low Risk)							
3	1 or 2	high	high	high	young	high	87.58
IF (Chest pain 1 OR 2) AND (Young) AND (Sugar high) OR (BP high AND Chol high) THEN (High Risk)							
4	4	low	medium	low	old	low	88.89
IF (Chest pain 4) AND (old) AND (Sugar low) OR (BP medium AND Chol low) THEN (Low Risk)							
5	3	medium	low	high	middle	medium	83.99
IF (Chest pain 3) AND (middle) AND (Sugar high) OR (BP low AND Chol medium) THEN (Medium Risk)							
6	2	high	high	low	middle	medium	82.35
IF (Chest pain 2) AND (middle) AND (Sugar low) AND (BP high AND Chol high) THEN (Medium Risk)							
7	1	high	medium	high	old	high	87.91
IF (Chest pain 1) AND (Old) AND (Sugar high) AND (BP medium OR Chol high) THEN (High Risk)							
8	2	medium	high	low	young	high	87.75
IF (Chest pain 2) AND (Young) AND (Sugar low) AND (BP high AND Chol medium) THEN (High Risk)							
9	1	low	medium	low	old	high	85.62
IF (Chest pain 1) AND (Old) AND (Sugar low) AND (BP medium OR Chol low) THEN (High Risk)							
10	4	low	low	low	old	medium	64.54
IF (Chest pain 4) AND (Old) AND (Sugar low) AND (BP low OR Chol low) THEN (Medium Risk)							
11	3 or 4	high	low	high	old	high	42.00
IF (Chest pain 3 OR 4) AND (Old) AND (Sugar high) OR (BP low AND Chol high) THEN (High Risk)							

Table 17: Proposed rules for the final Mamdani Model

7. BUILDING THE FINAL MAMDANI MODEL

7.1 Model Selection:

There are 3 commonly used Fuzzy Logic System models which are Mamdani, Tsukamoto and Sugeno model. These models are common in the sense that they take in a crisp value (categorical or numerical) and converts them into a fuzzy set domain. However, these 3 models have a different computational process in predicting the final out. A clear distinction between these 3 models is the process in which the final crisp output is generated from the fuzzy set.

The Mamdani model calculates the crisp output by evaluating the center of gravity / center of area under the fuzzy output membership curve within the limits that are defined by the fuzzy inputs and rules. According to [10], expert system applications such as medical diagnostics, are more inclined towards adopting Mamdani method as the model understands the bases of rule better as compared to Tsukamoto and Sugeno method.

Tsukamoto and Sugeno method on the other hand utilize the Weighted Average method to calculate the final crisp output. A comparison between the characteristics of these 3 models are depicted in Figure 16.

FIS Type	Fuzzification	Inference process	Defuzzification
Mamdani	Singletone	Min-max	Center of Gravity
Sugeno	Singleton	Min-product; Order 0	Weighted Average
Tsukamoto	Singletone	Min-max	Height Method

Figure 16: Comparison of the Mamdani, Tsukamoto and Sugeno Models[11]

The Fuzzy Logic System model selected for this study is the Mamdani Model due to the following reasons:

- The nature of the problem where the inputs and outputs are correlated and interpreted from an intuitive and in the form of a qualitative knowledge rule base in the form of IF (Antecedent) THEN (Consequent) rules. Due to this reason, Mamdani models are widely used for applications that especially require decision support requirements which is similar for the application of this problem in addition to the justification provided by [10]. [11][12]
- The final output value is categorical in nature and not numerical i.e., the problem is a classification problem and not a regression problem. For this reason, the Sugeno model was not considered as it is only applicable for regression-based problems.

7.2 Handling exceptions

Exceptions are defined as an event that occurs during the execution of a program that disrupts the normal flow of the code. Throughout the process of building the final Mamdani Fuzzy Model, from beginning stages of determining the best membership functions for each parameter, rule definition to the final model implementation, there were several occurrences where errors were faced or situations where the predicted outcome did not meet the expected outcome. These exceptions were dealt with by means of adding a block of code with conditions or redesigning the mode to handle the error. These errors and their respective mitigation steps are: -

1) Membership functions were defined in such a way to not have overlapping between subsequent class. A consequence to this was that there were particular data points at the border between two classes. These data points did not fall in any class and had zero memberships which is not only inaccurate but also presented an error when calculating the centroid of the curve. To mitigate this, a condition was set to take the smallest membership values from the class before and the class after based on the range of the entire dataset. An example of where this situation was face was when defining the membership function for input "Age". Details of the problem and solution were elaborated in Section 5.7.

2) When building the final model based on only the aggregate of the 9 rules selected from Section 6.2, it was noticed that the aggregated rule would result in a zero membership as most of the rules had AND conditions (i.e., the smallest value of membership was considered). Should any one of the dataset points have a value of zero membership for any of the 5 input parameters coinciding with the AND rule, the final membership value for that particular rule would be 0. If this condition is scaled up to all 9 rules the final aggregated membership will be 0 resulting in an error when calculating the centroid of the curve and is also not accurate. For example, considering an individual of the following input parameters and the following 3 rules considered: -

INPUTS				
Age: Old	Blood Pressure: High	Cholesterol: High	Sugar Level: Low	Chest Pain: CP1
RULE				**FINAL MEMBERSHIP**
Rule 1: IF (Middle OR Young) AND (chest pain 4) AND ((BP low OR BP medium) OR (Chol low OR Chol medium)) THEN (Low Risk)				0
Rule 2: IF (Chest pain 2) AND (middle) AND (Sugar low) AND (BP high AND Chol high) THEN (Medium Risk)				0
Rule 3: IF (Chest pain 1) AND (Old) AND (Sugar high) AND (BP medium OR Chol high) THEN (High Risk)				0

Though the individual should most likely be classified as Medium – High risk due to his/her age, blood pressure level, chest pain and cholesterol levels. However, the memberships under these 3 rules in addition with the AND condition directly reduces the patient's membership to zero when aggregated, significantly and inaccurately downplaying the individual's risk of heart disease.

Hence to mitigate this error, the 15 rules defined from Section 6.1 were also considered in the final model as these 15 rules covers and ensures that each datapoint / input parameter is assigned a value of membership no matter how low or high its value. Should its membership be reduced or increased due to multiple input-based rules set in Section 6.2, it is acceptable considering other a joint factor of inputs. However, this condition serves to cover datapoints that result in 0 memberships. The final aggregate rules consist of a total of 24 (15+9) rules.

3) A final condition that was set was regrouping the 5-level output membership to only a 3-level based output membership as explained in Section 5.8 due to the low accuracy when implementing with 5-tier output. In addition to this, as the defuzzied output calculated by the Mamdani model is in the form of a continuous numerical decimal number, the conditions detailed in Table 13 was also implemented to round / categorize the predicted output to their respective class of low, medium, or high risk.

7.3 Model Implementation

Upon finalizing the membership functions, the fuzzy system rules and the Fuzzy model, the final model was built using its building blocks. The memberships finalized as per Section 5.9 were built for each input and the output. It is important to note that the output membership function was redefined in the final model at early on stages unlike in Section 5. The membership function of the output was redefined as per Table 13 during membership evaluation and not after defuzzification.

The final 9 + 15 rules as established in Section 6 were then built and an aggregate of these rules were considered in terms of an OR condition. With an OR condition the maximum probability or membership for that particular dataset considering all the rules is taken as final, considering any one of the rules occurring. The membership / likelihood is maximized for this situation because in medical diagnosis, it is better for the model to be more sensitive in detecting high risk patients than to wrongly downplay the heart disease risk of the individual. If a person is flagged by the system further steps and precautions can be taken to confirm the diagnosis. [13]. The centroid of the area under the output membership curve (defined by rules) is then calculated to evaluate the defuzzied output. The defuzzied output is then categorized to its respective class. The dataset mentioned in Section 4 was used for testing purposes and the confusion matrix and accuracy was calculated for the 612 data points.

8. RESULTS AND DISCUSSION

The final model was able to predict the outcome with an accuracy of 87.58%. The confusion matrix produced by the model is displayed as below:

Predicted Output / Actual output	Low Risk	Medium Risk	High Risk
Low Risk	536	8	0
Medium Risk	52	0	0
High Risk	16	0	0

Table 18: Confusion Matrix of the Final Mamdani Model

From the Confusion matrix it can be seen a high source of accuracy can be attributed to the model's capability to predict low risk group. However, the model is still not as capable of predicting medium and high-risk patients. A reason for this can be explained by the high Low Risk group to High-Risk ratio in the raw data as depicted in Figure 14. Should more testing data on medium and high-risk individuals be acquired, the model could be tested more comprehensively to ascertain its performance with regards all class of patients. An effect of this imbalance in the raw data could affect the model's accuracy as it would be more biased towards diagnosing someone as healthy as opposed to being more sensitive in diagnosing high risk patients as the results show in Table 18.

The results and testing on the input data individually and all together demonstrated that each of the input had varying effects on the output. Initial observations on the accuracy of the simplified rules were that cholesterols did not significantly affect the outcome of heart disease as its accuracy did not amount to much when tested individually. This is concurrent with the idea proposed by [14] in recent medical journals. On the other hand, it was observed that increasing levels of fasting blood sugar and age played the strongest roles towards increasing the heart disease risk, followed by blood pressure and chest pain as it attributed to higher accuracies.

However, looking from another perspective, for some inputs, the high accuracy obtained may not be the sole criteria to judge the input's significance. Another parameter to look at is the variation in the output accuracy as the input parameter is varied. An example for this can be seen for blood sugar. Even with an accuracy of 88.89% on its own, blood sugar's significance was diminished when it was tested on the combined rules. When testing the 11 rules individually from Section 6.2 and when building the final model, it was observed that regardless of switching the fasting blood sugar between high or low, the output accuracy showed insignificant changes. The very little to no variation in the output accuracy when the input parameter is modulated could also indicate that the input parameter is not so significant in defining the final output. Blood sugar's high accuracy on its own can be explained by its simplified classification of only low and high whereby only 2 conditions can be set which are IF low blood sugar THEN healthy or, IF high blood sugar THEN sick 1, sick 2, sick 3 and sick 4. This results in the classification of the output data to only either one of 2 very largely covered but vague categories as there is less room for categorization. The result based on input data of blood sugar may present itself with high accuracy at first but after observing how it affects the output, it was hypothesized that its influence may not be as strong as

originally assumed. A solution to this would be to acquire data in the form of numerical blood sugar levels as compared to categorial data and redefine its memberships and rules.

Chest pain is considered one of the most influencing indicators of heart disease risk. An instance of its effect can be seen when developing combined rules in Section 6.2. When combined rule 11 associated Chest pain 3 OR 4 with high risk of heart disease, the accuracy of the model dropped significantly to 42%. However, looking at rule 2, 4 or 5 which associates Chest Pain 3 and 4 with Medium OR Low Risk, the accuracy increased to above 83%. Additionally, when rule 1 or 2 was associated with Medium to High-Risk categories, the accuracy is also above 83%. This pattern in the accuracy further reinforces the individual rules defined for chest pain in Section 6.1 as when the output conditions for Chest Pain 1 and 2 are replaced with Chest Pain 3 and 4, the accuracy dropped significantly.

Another observation that can be made was how the accuracy of the outcome increased as more inputs were evaluated together by the model. This can be seen for the instance of chest pain. When tested individually with simplified rules, chest pain was ranked the 2nd lowest among the 5 inputs in terms of output accuracy. However, when combined with other inputs in Section 6.2, the model produced more accurate outputs. Cholesterols and blood pressure also both had a moderate level of accuracy when tested individually which then increased when tested under combined rules. From the pattern observed from the data, it can be inferred that the effect of inputs on the outcome in a Fuzzy Logic System vary (dominate or diminish) as they are combined with other inputs and rules / conditions. More input parameters give the Fuzzy model more factors to consider, giving it better decision-making capabilities.

Looking at combined rules Rule 7 and Rule 9 in Section 6.2. Other input parameters have the same level of memberships except Cholesterol and Blood Sugar. It was already established that blood sugar had relatively low impact on the accuracy. Thus, these 2 rules are suitable for comparison of cholesterol under combined rule condition. From the accuracy it can be seen that when cholesterol levels were changed from low to high, the accuracy increased by a mere 2% which justifies the observation made in Section 6.1 and as explained above that cholesterol does not significantly affect the output accuracy.

Results from the membership functions definition for the inputs show that linear based functions (increasing, decreasing and triangle) represents the numerical input parameters less accurately with respect to physiological parameter readings as compared to sigmoid and gaussian functions. This was demonstrated in the testing phase of memberships for cholesterols and blood pressure, which ultimately adopted sigmoid and gaussian membership functions due to their higher accuracy.

9. CONCLUSION

Fuzzy Logic Systems are an area of Artificial Intelligence that produces a definite output, based on sets of inaccurate, unclear and fuzzy input conditions. Fuzzy logic operates based on imitating human based reasoning where outcomes are not always a definite YES or NO. Instead, it is based on various factors and possibilities within the range of YES or NO to predict a possible outcome.

This study aims to implement a Fuzzy Logic System based on easily available physiological inputs of individuals to ascertain his/her level of risk towards heart disease. While there are various risk factors for heart disease, a total of 5 risk factors were identified and selected as inputs for the Fuzzy Logic system of this project. The five inputs are Chest Pain, Cholesterol, Blood Pressure, Fasting Blood Sugar level and Age. The output is the heart disease risk level. Each input and output were linguistically defined, and its memberships were also defined.

A total of 15 rules were built for the Fuzzy system for each individual input as detailed in Section 6.1. Subsequently, 11 additional rules were defined with all inputs simultaneously considered by the model in Section 6.2. Out of the 11 rules, only 9 of the highest accuracy was selected for final model implementation. The final model implemented the Mamdani model due to its recognized application in the medical diagnosis area and because it handles rule-based systems much better.

There were several instances in the process of building the model where exceptions / errors occurred which are undesirable as it affects the accuracy and performance of the model. Hence, in order to mitigate these exceptions, certain conditions or steps were defined in the model implementation. These conditions are defined in Section 7.2 and include averaging values of memberships, regrouping output categories and considering more rules for the final rule aggregation in the model.

In order to handle exceptions as explained in Section 7.2, the Mamdani Fuzzy Logic System included the 15 rules initially defined in Section 6.1 in addition to the 9 final selected rules from Section 6.2. The final 24 rules encompassed all membership classes ensuring that no data points had zero membership, a situation that is not realistic. The final 24 rules implemented are displayed in Table 15.

Rule 1	IF (Chest pain 4), THEN (Low Risk)
Rule 2	IF (Chest pain 3), THEN (Low Risk)
Rule 3	IF (Chest pain 2), THEN (Medium OR High Risk)
Rule 4	IF (Chest Pain 1), THEN (Medium OR High Risk)
Rule 5	IF (Cholesterol low), THEN (Low Risk)
Rule 6	IF (Cholesterol medium), THEN (Low Risk)
Rule 7	IF (Cholesterol high), THEN (Medium OR High Risk)
Rule 8	IF (Blood Pressure Low) THEN (Low Risk)
Rule 9	IF (Blood Pressure Medium), THEN (Low Risk)
Rule 10	IF (Blood Pressure High), THEN (Medium OR High Risk)
Rule 11	IF (Blood Sugar Low), THEN (Low Risk)
Rule 12	IF (Blood Sugar High), THEN (High Risk)
Rule 13	IF (Age Young), THEN (Low Risk)
Rule 14	IF (Age Middle), THEN (Low Risk)
Rule 15	IF (Age Old), THEN (Medium OR High Risk)

Rule 16	IF (Age Middle OR Old) AND (Chest pain 1) AND (Blood Pressure High OR Cholesterol High) THEN (Medium OR High Risk)
Rule 17	IF (Age Middle OR Young) AND (Chest pain 4) AND ((Blood Pressure Low OR medium) OR (Cholesterol Low OR Medium)) THEN (Low Risk)
Rule 18	IF (Chest pain 1 OR 2) AND (Age Young) AND (Sugar high) OR (Blood Pressure High AND Cholesterol High) THEN (High Risk)
Rule 19	IF (Chest pain 4) AND (Age Old) AND (Sugar low) OR (Blood Pressure Medium AND Cholesterol Low) THEN (Low Risk)
Rule 20	IF (Chest pain 3) AND (Age Middle) AND (Sugar high) OR (Blood Pressure Low AND Cholesterol Medium) THEN (Medium Risk)
Rule 21	IF (Chest pain 2) AND (Age Middle) AND (Sugar low) AND (Blood Pressure high AND Cholesterol High) THEN (Medium Risk)
Rule 22	IF (Chest pain 1) AND (Age Old) AND (Sugar high) AND (Blood Pressure medium OR Cholesterol High) THEN (High Risk)
Rule 23	IF (Chest pain 2) AND (Age Young) AND (Sugar Low) AND (Blood Pressure High AND Cholesterol Medium) THEN (High Risk)
Rule 24	IF (Chest pain 1) AND (Age Old) AND (Sugar Low) AND (Blood Pressure Medium OR Cholesterol Low) THEN (High Risk)

Table 19: Table of the final 24 rules used in the Model

Several key areas of improvement to the model can be made in terms of the inputs and rules considered. With more inputs and rules, more combinations of scenarios can be covered, and it would make the model more accurate in evaluating the heart disease risk level. Another area of improvement is in terms of the balance of output class for dataset used for testing. As displayed and explained in Figure 14 and Section 5.8, a majority of the data points available are skewed towards the Low-Risk end with significantly lesser data points for Medium and High-Risk Groups. This is also one of the reasons why the output was regrouped from 5 to 3 classes. Having an imbalanced outcome group in the datasets used for analysis and testing would make the predicted outcome by the model more biased towards the Low-Risk end which would ultimately affect the accuracy. There is also no proper evaluation to determine the model's accuracy in diagnosing Medium and High-Risk groups. Hence, more datapoints can be added to testing the model and with a good balance between each output class. Additional analysis on the observations or findings and different methods of evaluating the weightage / influence of each input can also be carried out to reinforce the rules and define more accurate linguistic definitions and memberships, which will ultimately improve the accuracy of the model. Lastly, it will be preferable to have inputs that are numerical in nature instead of categorial. For example, should the input data of diabetes be in the form of readings in mg/dl, it would produce more accurate memberships. Having numerical data gives more flexibility in defining fuzzy memberships and building the Fuzzy model. These improvements can certainly be implemented given additional time and resources.

In overall, the Fuzzy Logic System for Heart Disease diagnosis was built and implemented using Mamdani Model. The final model managed to achieve an overall accuracy of 87.58%.

10. REFERENCES

[1] UCI Machine Learning Repository (2021) *Heart Disease Data Set*, *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ml/datasets/heart+disease (Accessed: 24 February 2021).

[2] Adeli, A. and Neshat, M. (2010) 'A Fuzzy Expert System for heart disease diagnosis', Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, IMECS 2010, (March), pp. 134–139.

[3] Sikchi, Smita, Sikchi, Sushil and M. S., A. (2013) 'Generic Medical Fuzzy Expert System for Diagnosis of Cardiac Diseases Smita Sushil Sikchi', International Journal of Computer Applications, 66(13), pp. 975–8887. Available at: https://pdfs.semanticscholar.org/2364/878203c25949e9c703cef3f9faba76039733.pdf.

[4] Allahverdi, N., Torun, S. and Saritas, I. (2007) 'Design of a Fuzzy Expert System for determination of coronary heart disease risk', ACM International Conference Proceeding Series, 285, pp. 1–8. doi: 10.1145/1330598.1330638.

[5] Bhuiya, F. A., Pitts, S. R. and McCaig, L. F. (2010) 'Emergency department visits for chest pain and abdominal pain: United States, 1999-2008.', NCHS data brief, (43), pp. 1–8.

[6] Oad, K. K., Dezhi, X. and Butt, P. K. (2014) 'Software & Data Engineering A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease', 14(3).

[7] Barman, M. and Chaudhury, J. P. (2013) 'A Framework for Selection of Membership Function Using Fuzzy Rule Base System for the Diagnosis of Heart Disease', International Journal of Information Technology and Computer Science, 5(11), pp. 62–70. doi: 10.5815/ijitcs.2013.11.07.

[8] Akhoondi, R., Hosseini, R. and Branch, S. (2016) 'c r v i h o e f c f', 7(2), pp. 101–114.

[9] Jain, P. and Kaur, A. (2019) 'A fuzzy expert system for coronary artery disease diagnosis', ACM International Conference Proceeding Series. doi: 10.1145/3339311.3339358.

[10] MathWorks (2021) *Mamdani and Sugeno Fuzzy Inference Systems*, *MathWorks*. Available at: https://www.mathworks.com/help/fuzzy/types-of-fuzzy-inference-systems.html (Accessed: 23 February 2021).

[11] Sari, W. E., Wahyunggoro, O. and Fauziati, S. (2016) 'A comparative study on fuzzy Mamdani-Sugeno-Tsukamoto for the childhood tuberculosis diagnosis', *AIP Conference Proceedings*, 1755(July 2016). doi: 10.1063/1.4958498.

[12] Wolkenhauer, O. (2003) 'Fuzzy Systems and Identification', *Data Engineering*, (May), pp. 109–128. doi: 10.1002/0471224340.ch5.

[13] 'A Comparative Analysis of Feature Extraction Techniques for EEG Signals from Alzheimer patients' (2012), 2012.

[14] Soliman, G. A. (2018) 'Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease', *Nutrients*. MDPI, 10(6), p. 780. doi: 10.3390/nu10060780.

11. APPENDICES

NO	ATTACHMENT NAME	ATTACHMENT DESCRIPTION	FILE
1	chest_pain.py	The python codes for chest pain membership definition	
2	cholesterol.py	The python codes for cholesterol membership definition	
3	blood_pressure.py	The python codes for blood pressure membership definition	
4	blood_sugar.py	The python codes for blood sugar membership definition	
5	age.py	The python codes for age membership definition	
6	final_model.py	The python codes for the final Mamdani Model	
7	Memberships.py	The python codes consisting of functions used for memberships and in the final model	
8	CLEVELAND.csv	3 .csv files of the raw dataset.	
	HUNGARIAN.csv		
	SWITZERLAND.csv		