symmetry

MDPI

Article

The Combinations of Fuzzy Membership Functions on Discretization in the Decision Tree-ID3 to Predict Degenerative Disease Status

Endang Sri Kresnawati [1,2], Bambang Suprihatin [2] and Yulia Resti [2,*]

[1] Doctoral Study Program, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Padang Selasa Bukit Besar, Palembang 30139, Sumatera Selatan, Indonesia; eskresnawati@unsri.ac.id

[2] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih, Km.32, Indralaya 30062, Sumatera Selatan, Indonesia; bambangs@unsri.ac.id

* Correspondence: yulia_resti@mipa.unsri.ac.id

Abstract: Degenerative diseases are one of the leading causes of chronic disability on a global scale, significantly affecting the quality of life of sufferers. These diseases also burden the health care system and individuals financially. The implementation of preventive strategies can be postponed until an accurate prediction of the disease status can be achieved. Degenerative diseases that are the leading cause of death in many countries are coronary heart disease (CHD), while diabetes mellitus disease (DMD) increases the risk of CHD. Most of the predictor variables from the dataset to predict the status of both diseases are continuous. However, not all prediction methods, including the Decision Tree Iterative Dichotomiser3 (DTID3) method, can process continuous data. This work aims to predict the status of both degenerative diseases, CHD and DM, using the DTID3 method with continuous type predictor variables transformed using discretization with the concept of set membership. Seven prediction models using the DTID3 method are proposed to predict the status of each degenerative disease. One DTID3 model uses the concept of crisp set membership, and six DTID3 models use the concept of fuzzy set membership (FDTID3). Each prediction model of FDTID3 represents one combination of fuzzy membership functions in discretizing continuous predictor variables, and one combination consists of three membership functions. The performance of the proposed FDTID3 model depends on the fuzzy membership functions used. The hypothesis that the performance of the seven proposed models differs at least in one metric and that the performance of the FDTID3 models is higher than the DTID3 model discretized using the concept of crisp sets has been proven.

Keywords: DTID3; degenerative disease status; discretization; fuzzy membership function; model performance

check for updates

Citation: Kresnawati, E.S.; Suprihatin, B.; Resti, Y. The Combinations of Fuzzy Membership Functions on Discretization in the Decision Tree-ID3 to Predict Degenerative Disease Status. *Symmetry* **2024**, *16*, 1560. https://doi.org/10.3390/sym16121560

Academic Editor: Wentao Mao

Received: 18 May 2024
Revised: 15 November 2024
Accepted: 18 November 2024
Published: 21 November 2024

1. Introduction

Continuous-type variables can be found in many real-life cases, especially datasets in the medical field [1,2]. This dataset is useful for predicting disease status, including degenerative [3,4]. Degenerative diseases occur due to a slow decline in the function of the body's organs and tissues and can attack the nerves, spine, joints, and brain [5]. These diseases tend to worsen over time and have an impact on the sufferer's quality of life [6–8]. Degenerative diseases are chronic because they are not contagious, develop slowly, and are long lasting [9]. Additionally, these illnesses rank as the leading cause of chronic disability on a global scale. Degenerative diseases impact over 30% of the worldwide populace, with the allocation of 70% of public health resources towards their treatment. Furthermore, degenerative diseases impose a substantial financial strain on health care systems and individuals alike [10].

Numerous fatalities are attributed to coronary heart disease, a degenerative condition caused by the blockage or narrowing of the coronary arteries due to fat deposition.

Moreover, it has emerged as the leading cause of mortality on a global scale [11]. With a substantial increase from the 12.1 million heart disease-related fatalities documented in 1990 to the 20.5 million coronary heart disease-related deaths recorded in 2021, this number represents an approximate 33.33 percent of all global deaths [12]. Untimely mortality is predominantly attributed to cardiovascular disease. One hundred and forty-six countries reported male fatalities, and ninety-eight countries reported female fatalities [12].

According to the Global Health Estimates published by the World Health Organization (WHO) in 2019, diabetes was responsible for approximately 1.5 million deaths [13]. Diabetes mellitus is a degenerative disorder that has the potential to give rise to numerous severe ailments [14]. The number of individuals with diabetes is projected to reach 313 million by 2040, according to the WHO [15]. It is anticipated that this trend will continue to worsen over time. Regarding economic strain, diabetes treatment will double in price from $13,700 annually by 2030 [16]. Prevention strategies can be implemented at an earlier stage if accurate status prediction of diabetes mellitus and coronary heart diseases are feasible and substantial cost reductions can be achieved in the treatment of both diseases at [17–19].

Several studies have made predictions using various methods regarding the presence of both degenerative diseases and proposed several approaches to improve their performance, such as ensemble techniques [20,21], class balancing on the dataset [19,22], feature scaling [23], selecting significant variables [24–26], handling missing value [27,28], or perform preprocessing before the data are predicted, such as transforming the data to a particular type because the prediction method requires the predictor variable to be of a specific type [29].

The Decision Tree Iterative Dichotomiser3 (DTID3) is a nonparametric prediction method that often provides satisfactory prediction performance in many cases, but the predictor variables in this method must be categorical. In the case of numeric predictor variables, they need to be transformed first into a categorical type, and one of the transformation techniques is crisp discretization, known as discretization (only) [30]. Discretization can also broaden knowledge regarding continuous data types [31] and enhance model performance [32,33]. Nevertheless, ambiguity may result from discretization [34]. The fuzzy set membership concept can be employed to rectify ambiguity in the context of discretization and is known as fuzzy discretization [35–37].

Using fuzzy discretization in several prediction methods has been shown to enhance prediction performance, such as the naïve Bayes (NB) method [35,36,38], neural network-radial basis function (NNRBF) [39], multilayer perceptron (MLP) [31], and the DTID3 [34,40]. Several different combinations of membership functions have been applied in these studies, as well as final membership selection rules. Unfortunately, so far there are no guidelines for selecting a combination of fuzzy membership functions in discretizing a predictor variable as well as the final membership selection rules, so trial and error [40,41] is still the best solution in determining the performance of a prediction model [35]. Regarding the amount of discretization of predictor variables, referring refers to expert justification [33,38] or prior knowledge [35].

Differences in fuzzy membership function combinations can affect the prediction model's performance. Research [35] classifies corn plant diseases and pests into seven classes using the naïve Bayes method. It proposes six combinations of fuzzy membership functions consisting of four combinations of fellow linear functions (all triangular, all trapezoidal, decreasing–increasing linear and triangular, decreasing–increasing linear and trapezoidal) and two combinations of nonlinear functions and linear functions (decreasing–increasing sigmoid and triangular, decreasing–increasing sigmoid and trapezoidal). The study obtained the best model performance using fellow linear functions: decreasing–increasing linear and triangular. However, each of the other five combinations had significantly different performance and experienced increased performance from the model using crisp discretization. Research [34] which classifies corn plant diseases and pests into six classes using the DTID3 method and discretizes predictor variables by combining nonlinear and linear fuzzy membership functions (decreasing–increasing sigmoid and triangular) also shows an increase in the performance of the model using crisp discretization. Likewise, research [42] also shows an increase in the performance of the original method (without fuzzy discretization). The study classifies can types using the naïve Bayes method and discretizes predictor variables with a combination of fellow linear fuzzy membership functions, namely decreasing–increasing linear and triangular. All three studies use the final membership selection rules of maximum value. The performance improvement of each study had accuracy at 0.7, recall at 3.95 [35], accuracy at 3.23, recall at 11.8 [34], and accuracy at 34.93, recall at 35.08 [42].

Another selection rule is the arithmetic mean, as in the research of [36,43], which predicts CHD status. Both use the naïve Bayes prediction method and uniform fuzzy membership functions for discretization, namely, all triangular and all trapezoidal, with performance improvements of accuracy at 4.03, recall at 4.05 in [36], and accuracy at 7.5 in [38].

In addition to the two final membership selection rules, several studies use defuzzification with a specific method such as centroid of area (COA) [39], mean of maximum (MoM), [37], or centroid of area (COA) [43]. Defuzzification is part of the fuzzy rule in fuzzy logic. In prediction tasks, the fuzzy logic method can stand alone, as in the research [43], or be integrated with other prediction methods, such as neural networks [37,39]. The study [39], which uses a uniform fuzzy membership function (all trapezoidal) to predict breast cancer, increased accuracy to 3.72 and recall to 10.96. However, the research of [37], which also used a uniform fuzzy membership function (all triangular), obtained unexpected performance in predicting DM status, where the performance of the initial model did not increase, and even decreased (accuracy at 18.35 and recall at 23.97). This fact is suspected due to the numbers of discretization in each predictor variable that does not refer to expert justification even though the data used is medical data, namely the DM dataset. In addition, the neural network prediction method does not require a transformation of the predictor variable into categorical because it can directly process data of the fuzzy discretization numeric type. In the CHD dataset, the research of [36] also subjectively discretized the predictor variables into two categories, not based on expert justification, which can vary for each variable. From all these studies, a combination of nonlinear fuzzy membership functions has not been found in discretizing the predictor variables. Specifically, for research discussing disease status prediction for CHD and DM, no research has been found that uses maximum value as the final membership selection rule or different combinations of fuzzy membership functions.

For these reasons, the main contribution of this study is to build prediction models of the status of the two degenerative diseases, CHD and DM, using the DTID3 method by discretizing continuous-type predictor variables using the concept of set membership. One DTID3 model was built using the concept of crisp set membership, and six DTID3 models whose variables were discretized using the concept of fuzzy set membership (FDTID3). The six combinations consist of two combinations of fellow linear functions (decreasing–increasing linear and triangular, decreasing–increasing linear and trapezoidal), two combinations of linear and nonlinear functions (decreasing–increasing sigmoid and triangular, decreasing–increasing sigmoid and trapezoidal), and two combinations of fellow nonlinear functions (decreasing–increasing sigmoid and beta, decreasing–increasing sigmoid and pi). The final membership selection rules used are maximum value. This study also hypothesizes that the performance of the seven models built differs in at least one metric. The performance of the six FDTID3 models is higher than that of the DTID3 model built using the crisp set membership concept.

2. Materials and Methods

2.1. Research Dataset

The dataset used in this research is the degenerative diseases dataset, CHD, and DM. Both are chosen for this primary reason: both data have a majority of predictor variables of continuous type (CHD). Even in DM, all predictor variables are the continuous type.

The second reason is to show that missing data of less than 2% do not affect the prediction model's performance using the proposed DTID3 method. In addition, the proposed DTID3 method can still perform well without giving treatment, such as discarding or imputing it with a particular value.

2.1.1. Coronary Heart Disease (CHD) Dataset

The CHD dataset was obtained free of charge via https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset, accessed on 5 November 2020 [44]. This dataset was created by Robert Detrano, M.D., Ph.D., in July 1988 based on the Cleveland Databases. The predictor variable of the Cleveland Databases CHD is presented in Table 1.

Table 1. The predictor variable of the CHD dataset.

Variable	Description	Type	Information
Age	age in years	Continuous	29–77 years
Sex	sex	Categoric	0: male 1: female
CP	chest pain type	Categoric	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Trestbps	resting blood pressure	Continuous	94–200 mmHg
Chol	serum cholesterol	Continuous	126–564 mg/dL
FBS	fasting blood sugar > 120 mg/dL	Categoric	0: false 1: true
Restecg	resting electrocardiographic results	Categoric	0: normal 1: having ST-T wave abnormal (>0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalach	maximum heart rate achieved	Continuous	71–202 bpm
Exang	exercise-induced angina	Categoric	0: no 1: yes
Oldpeak	ST depression induced by exercise relative to rest	Continuous	0–6.2 mV
Slope	the slope of the peak exercise ST segment	Categoric	1: upsloping 2: flat 3: down sloping
Ca	number of significant vessels colored by fluoroscopy	Discrete	0–3
Thal	thalassemia (types of blood disorder)	Categoric	3: normal 6: fixed defect 7: reversible defect

The CHD dataset has a size of 303 observations distributed into the Yes class (patients with the status of having CHD) at 45.87% and the No class (patients with the status of not having CHD) at 54.13%. Five of the thirteen predictor variables are continuous and can be transformed into categorical variables using crisp or fuzzy discretization. The number of categories in each discretized variable refers to the expert justification related to CHD as in [33]. Further exploration is needed, considering that almost all predictor variables except the Age variable have zero observation values.

2.1.2. Diabetes Mellitus Dataset (DMD)

The DMD was obtained at no cost through https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, accessed on 9 October 2020 [45]. The participants were

768 women aged 21 to 81 years who had undergone individual diagnostic measurements and were classified into class Yes (participants who had DM) at 65.1% and class No (participants who did not have DM). The predictor variables of the DMD are presented in Table 2.

Table 2. The predictor variable of the DMD.

Variable	Description	Information
Glucose (ratio)	Plasma glucose concentration 2 h in an oral glucose tolerance test	0–199 mg/dL
Blood Pressure	Diastolic blood pressure (blood pressure when the heart relaxes)	0–122 mmHg
Skin Thickness	Triceps skin fold thickness	0–99 mm
Insulin	2-Hour serum insulin	0–846 µ/mL
BMI	Body mass index (an approximate of total body fat)	0–67.1 kg/m^2
Diabetes Pedigree Function	a function that scores the probability of diabetes based on family history	0.08–2.42
Age	Age in years	21–81 years
Pregnancies	Number of times pregnant	0–17 times

In the DMD, all eight predictor variables are of continuous type, so all of them can be transformed into categorical variables using either crisp or fuzzy discretization. The number of categories in each discretized variable refers to expert justification related to DM as in [46]. Further exploration is needed, considering that almost all predictor variables except the Age variable have zero observation values.

2.2. Research Method

The steps of the proposed method in this study are given in a flowchart, as shown in Figure 1.

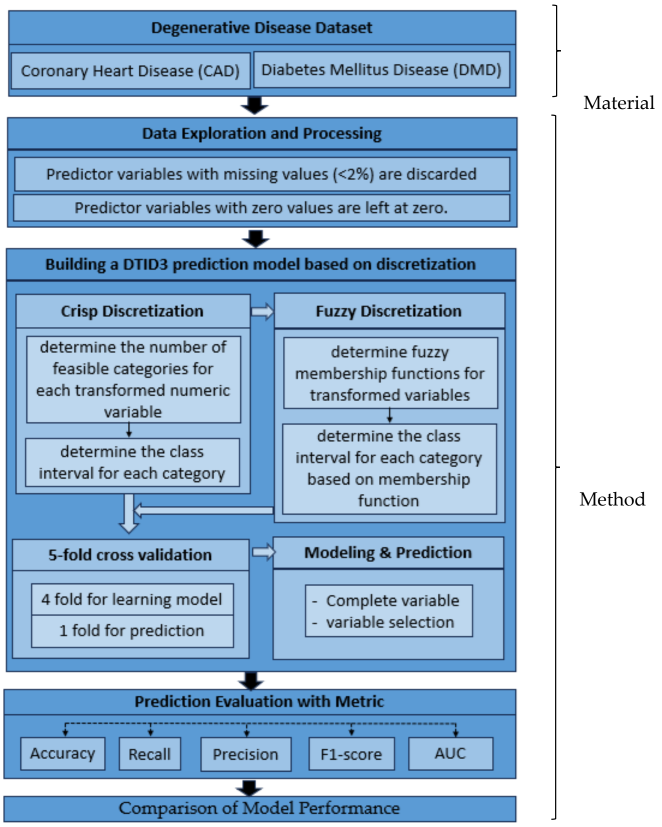

Degenerative Disease Dataset
Coronary Heart Disease (CAD)
Diabetes Mellitus Disease (DMD)
Material
Data Exploration and Processing
Predictor variables with missing values (<2%) are discarded
Predictor variables with zero values are left at zero.
Building a DTID3 prediction model based on discretization
Crisp Discretization
determine the number of feasible categories for each transformed numeric variable
determine the class interval for each category
Fuzzy Discretization
determine fuzzy membership functions for transformed variables
determine the class interval for each category based on membership function
5-fold cross validation
4 fold for learning model
1 fold for prediction
Modeling & Prediction
- Complete variable
- variable selection
Method
Prediction Evaluation with Metric
Accuracy
Recall
Precision
F1-score
AUC
Comparison of Model Performance

Figure 1. The flowchart of the proposed method.

2.2.1. Data Exploration and Processing

In the first step, this work explores the research variable, ignores the missing data, which is less than 2%, and leaves the zero data as zero without discarding them or imputing them with the mean, median, or mode. The missing data of less than 2% do not significantly affect the prediction of disease status. Zero data differ from missing data, and the DTID3 method can handle zero data so that they do not need to be imputed.

2.2.2. Building the DTID3 Prediction Model Based on Discretization

The DTID3 is a nonparametric classification method. This method does not use statistical assumptions such as variable independence, multicollinearity, or the presence of interaction effects between variables. The method uses the concept of entropy in making decisions presented in the form of nodes that form a tree-like flow diagram. Decisions at each node are taken based on the information gained from each categorical variable, in each iteration. Information gain is defined in Equation (9) [33].

$$Information\ Gain(S,\ X) = Entropy\ (S) - \sum_{c=1}^{k_X} \frac{|S_c|}{|S|} Entropy\ (S_c) \tag{1}$$

The S and S_c represent the total number of patients and the total number of patients in the c-category of the predictor variable X, the $Entropy\ (S)$ is defined as

$$Entropy\ (S) = \sum_{i=1}^{k_s} -P_i\ log_2 P_i \tag{2}$$

For P_i being the prior probability of the i-th class [33,46].

This method cannot directly process continuous predictor variables, it must first be transformed into a categorical type. This proposed the discretization to handle this problem. The $Entropy\ (S_c)$ for each crisp and fuzzy discretization is given by

$$Entropy\ (S_c) = \sum_{c=1}^{k_X} -P_c\ log_2 P_c \tag{3}$$

$$Entropy\ (S_c) = \sum_{c=1}^{k_X} -P_{cf}\ log_2 P_{cf} \tag{4}$$

P_c and P_{cf} being the prior probabilities of the c-th category of the predictor variable X, which are discretized using the crisp and the fuzzy sets, respectively. P_{cf} is obtained using

$$P_{cf} = \sum_{f=1}^{F} P_c\ \mu_{cf} \tag{5}$$

The μ_{cf} represent the fuzzy membership function for the c-th category of the predictor variable X [33,34].

Under the research objectives, seven prediction models using the DTID3 method are built to predict the status of each degenerative disease. One DTID3 model was built using the crisp set membership concept, and six DTID3 models were built using the fuzzy set membership concept (FDTID3). Each prediction model of FDTID3 represents a combination of fuzzy membership functions in discretizing continuous predictor variables, and one combination consists of three membership functions with the same pattern displays symmetry. The first category applies a decreasing pattern to each variable in each combination, including linear and nonlinear functions. The second group uses fuzzy memberships with symmetrical curve forms, such as triangular, trapezoidal, pi, and beta memberships. The third category of discretization employs fuzzy membership with increasing patterns, both linear and nonlinear functions. The six FDTID3 combinations consist of two combinations that are linear functions, FDTID3-1 (linear decreasing–increasing and triangular) and FDTID3-2 (linear decreasing–increasing and trapezoidal), then two combinations of linear and nonlinear functions, FDTID3-3 (sigmoid decreasing–increasing and triangular) and FDTID3-4 (sigmoid decreasing–increasing and trapezoidal), and finally two combinations that are nonlinear functions, FDTID3-5 (sigmoid decreasing–increasing and beta) and FDTID3-6 (sigmoid decreasing–increasing and pi). The final membership selection rules used are maximum value. This study also hypothesizes that the performance of the six models and the DTID3 model with crisp discretization differs at least in one metric.

Specifically, discretization is a preprocessing method that converts continuous data into categorical data [30]. This method is indispensable in specific statistical machine-learning techniques that necessitate categorical predictor variables, including the DTID3 method [33,34,40], multinomial naive Bayes method [36,46–48] and others. Crisp sets or fuzzy sets may be implemented for discretization. A crisp set is a collection of elements with a membership degree of 1. If it is not an element of a set, then its membership is zero. In contrast to the crisp set, the fuzzy set is a collection of elements with a degree of membership in interval [0, 1]. The degree of membership is given by a function called the fuzzy membership function [49].

Crisp discretization, which employs crisp sets, is characterized by mutually exclusive category intervals. Crisp discretization can be formed based on expert justification [33,46] prior information in specific fields of science [46], or using a formula as in (6) [34,35]. Conversely, fuzzy discretization, which employs fuzzy sets, allows for overlapping intervals for categories. These overlapping intervals are appropriate when a variable's categorization is defined flexibly by antecedent information. The fuzzy discretization can be formed based on crisp discretization [34]. ${X_d}^o$ and X_d are the d-th initial predictor variable and d-th, discretized predictor variables, respectively.

$$X_d = {X_d}^o + \text{Range}({X_d}^o) \tag{6}$$

Furthermore, the interval class boundaries for fuzzy discretization are obtained using fuzzy membership function parameters. Fuzzy membership functions can be either linear or nonlinear functions.

A linear function is a function with either one or two variables without exponents. This function represents a straight line on the coordinate plane. The linear fuzzy membership functions in this research are represented by Equations (7)–(10), where each represents membership functions called linear ascending, linear descending, triangular, and trapezoidal.

For the decreasing linear fuzzy membership function in (7), a is the smallest domain element with a membership degree of 1, and b is the most prominent domain element with a membership degree of 0 [50].

$$\mu(x;a,b) = \begin{cases} 1 & ; \quad x \le a \\ \frac{b-x}{b-a} & ; \quad a \le x \le b \\ 0 & ; \quad x \ge b \end{cases} \tag{7}$$

In contrast to the decreasing linear, in the increasing linear fuzzy membership function in (8), a is the smallest domain element with a membership degree of 0, and b is the most prominent domain element with a membership degree of 1 [50].

$$\mu(x;a,b) = \begin{cases} 0 & ; \quad x \le a \\ \frac{x-a}{b-a} & ; \quad a \le x \le b \\ 1 & ; \quad x \ge b \end{cases} \tag{8}$$

Equation (9) shows the triangular fuzzy membership function with a being the smallest domain element that has a membership degree of 0, b being a domain element that has a membership degree of 1, and c being the most prominent domain element that has a membership degree of 0 [51].

$$\mu(x;a,b,c) = \begin{cases} 0 & ; \quad x \le a \\ \frac{x-a}{b-a} & ; \quad a \le x \le b \\ \frac{c-x}{c-b} & ; \quad b \le x \le c \\ 0 & ; \quad x \ge c \end{cases} \tag{9}$$

For the trapezoidal fuzzy membership function in (10), which has four parameters, a is the smallest domain element that has a membership degree of 0, b is a domain element that has a membership degree of 1, c is a domain element that has a membership degree of 1, and d is the most prominent domain element that has a membership degree of 0 [51].

$$\mu(x;a,b,c,d) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \frac{d-x}{d-c} & c \le x \le d \\ 0 & x \ge d \end{cases} \tag{10}$$

In some cases, the fuzzy membership is also related to the nonlinear function. A linear equation is used to represent a straight line in a graph, whereas nonlinear equations are used to represent curves. Equations (11)–(14) show each nonlinear fuzzy membership function used in this research, namely decreasing sigmoid, increasing sigmoid, pi, and beta for random variable X.

$$\mu(x;\alpha,\beta,\gamma) = \begin{cases} 1 & ; \quad x \le \alpha \\ 1 - 2\left(\frac{x-\alpha}{\beta-\alpha}\right)^2 & ; \quad \alpha \le x \le \beta \\ 2\left(\frac{\gamma-x}{\gamma-\beta}\right)^2 & ; \quad \beta \le x \le \gamma \\ 0 & ; \quad x \ge \gamma \end{cases} \tag{11}$$

In the decreasing sigmoid fuzzy membership function in (6), α is the smallest domain element, which has a membership degree of 1, β is the domain element as an inflection point, which has a membership degree of 0.5, and γ is the most prominent domain element which has a membership degree of 0 [52].

$$\mu(x;\alpha,\beta,\gamma) = \begin{cases} 0 & ; \quad x \le \alpha \\ 2\left(\frac{x-\alpha}{\beta-\alpha}\right)^2 & ; \quad \alpha \le x \le \beta \\ 1 - 2\left(\frac{\gamma-x}{\gamma-\beta}\right)^2 & ; \quad \beta \le x \le \gamma \\ 1 & ; \quad x \ge \gamma \end{cases} \tag{12}$$

For the increasing sigmoid fuzzy membership function in (7), α is the smallest domain element, which has a membership degree of 0, β is the domain element as an inflection point, which has a membership degree of 0.5, and γ is the most prominent domain element which has a membership degree of 1 [52].

$$\mu(x;\gamma,\beta) = \begin{cases} 0 & ; \quad x \leq \gamma - \beta \\ 2\left(\frac{x-\gamma+\beta}{\beta}\right)^2 & ; \quad \gamma - \beta \leq x \leq \gamma - \frac{\beta}{2} \\ 1 - 2\left(\frac{\gamma-x}{\beta}\right)^2 & ; \quad \gamma - \frac{\beta}{2} \leq x \leq \gamma \\ 1 & ; \quad x = \gamma \\ 1 - 2\left(\frac{x-\gamma}{\beta}\right)^2 & ; \quad \gamma \leq x \leq \gamma + \frac{\beta}{2} \\ 2\left(\frac{\gamma+\beta-x}{\beta}\right)^2 & ; \quad \gamma + \frac{\beta}{2} \leq x \leq \gamma + \beta \\ 0 & ; \quad x \geq \gamma + \beta \end{cases} \tag{13}$$

Equation (8) shows the pi fuzzy membership function with γ being the center of the curve, a domain element with a membership value of 1, and β being the curve's width, from the curve's center to the curve's end [52].

$$\mu(x;\gamma,\beta) = \mu_A(x) = \frac{1}{1 + \left(\frac{x-\gamma}{\beta}\right)^2} \qquad ; \qquad \gamma - \beta \leq x \leq \gamma + \beta \tag{14}$$

The beta fuzzy membership function presented in Equation (9) has a parameter γ, a domain element with a membership value of 1, and is also the center of the curve. In contrast, the parameter β is the midpoint of half the width of the pi curve [52].

Five-Fold Cross Validation

This work uses the k-fold cross-validation technique in measuring model performance for unseen data where k = 5. The dataset is randomly divided into five-folds of equivalent size, with one-fold designated as testing data and the remaining four designated as learning data. There are five iterations with different training and testing data compositions. Each iteration contains training and test data with a ratio of 80:20. An overview of 5-fold cross-validation can be seen in Figure 2.

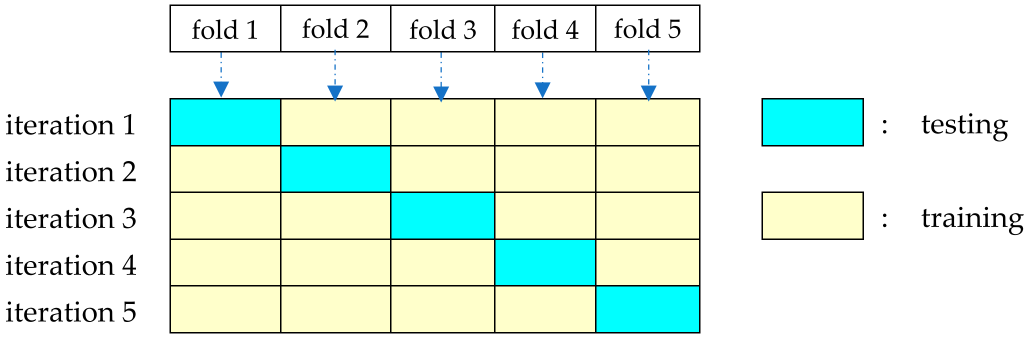


Figure 2. Five-fold Cross-validation.

The final model performance is the average of the five data compositions [53]. The 5-fold is one of the k-fold techniques that is less biased [54]. This method does not reduce the amount of data used for the learning model, can assess the quality of the fitted model and the stability of its parameters, and can also avoid overfitting.

2.2.3. Prediction Evaluation Metric

The best model among all proposed models is evaluated using performance measures. The higher the value of the performance measures metrics indicates the better model. The evaluation of the degenerative prediction models' performance in this study is conducted by calculating metrics of binary class: Accuracy, Recall, Precision, F1-scores, and AUC from (15) to (19) [55].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

$$AUC = \frac{TP}{2(TP + FN)} + \frac{TN}{2(TN + FP)} \tag{19}$$

TP is a True Positive prediction, FP is a False Positive prediction, TN is a True Negative prediction, and FN is a False Negative prediction. The accuracy metric is employed to determine the precision of a classification model. The Recall metric is employed to quantify the genuine positive ratio. Precision is calculated as the proportion of accurately anticipated positive status classes concerning the overall number of positively predicted classes. The F-1 Score is calculated by averaging the harmonic products of Precision and Recall. The area under the curve (AUC) is the probability of accurately predicting an observation. This metric indicates the extent to which the model can distinguish between the two classes, irrespective of the threshold selected. The model's quality is elevated as it increases. Its value falls from 1 to 0 [55].

3. Results and Discussion

3.1. Coronary Heart Disease Dataset

3.1.1. Dataset Exploration and Preprocessing

Two of the thirteen predictor variables in the dataset have missing values, namely the Ca variable (0.66%) and the Thal variable (1.32%). Missing values in both variables involve six patients. This work ignores the missing data that is less than 2% since it does not significantly affect the prediction of CHD status, so the total observations become 297 patients with a composition of 46.13% in the Yes class and 53.87% in the No class. Five of the 13 predictor variables in the dataset are of the continuous type, one is of the discrete type, and the rest are of the categorical type. The histogram of each continuous predictor variable and the bar plot of the discrete variable are given in Figure 3.

None of the five continuous predictor variables showed a normal distribution. This fact is supported by several statistical test results, such as those of Kolmogorov–Smirnov, Anderson Darling, and Cramer von Mises, indicated by a p-value less than the 5% significance level. The majority of the distributions of these variables are skewed to the right, but the Age and Thalach variables tend to be skewed to the left. The bar plot of Ca shows that one has the highest frequency, and the number continues to decrease as the frequency increases.

Table 3 displays a summary of continuous variables in each patient's status. All of the variables in both classes have value intervals that overlap. Not all variables in class Yes have a higher value interval than class No, as does the mean value. The maximum values of the variables of Cholesterol and Thalach in the Yes class are lower than the No class, and the minimum value of the Thalach in the No class is higher than the Yes class, as is the mean value.

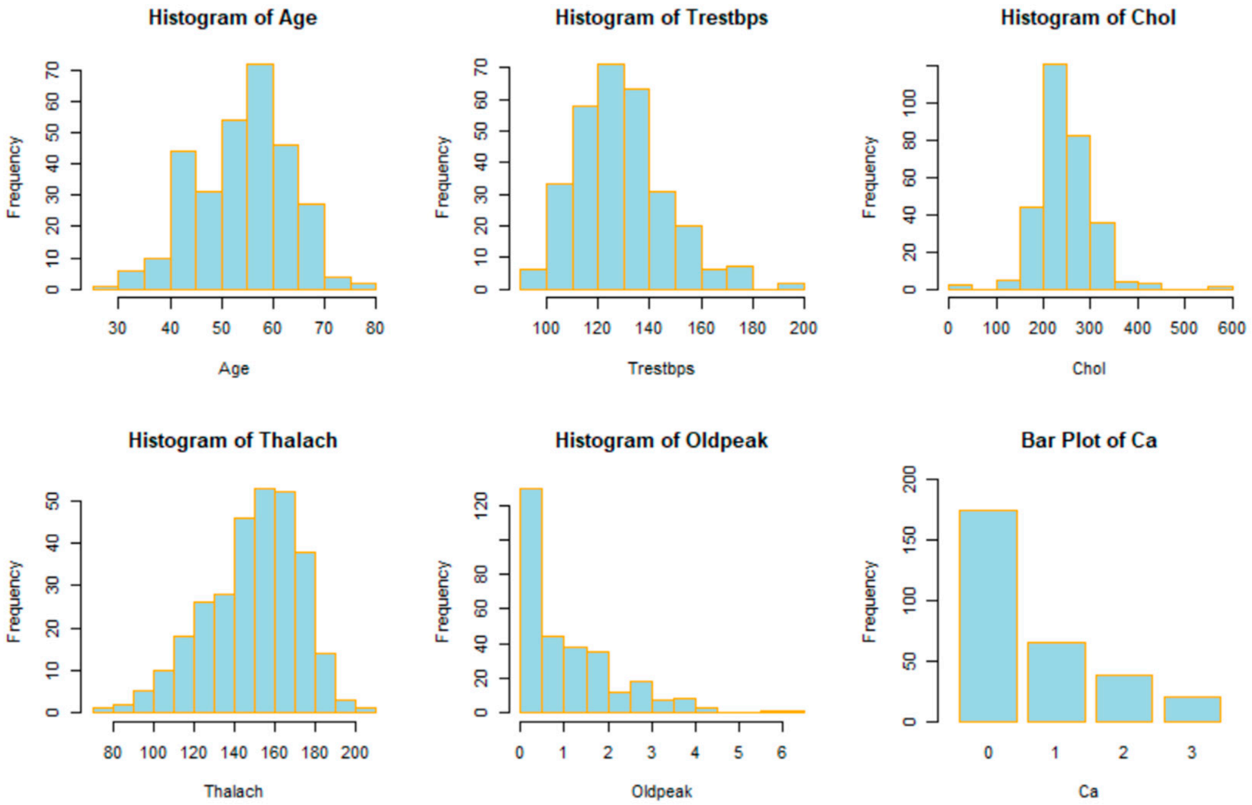


Figure 3. Histogram and bar plot of predictor variable in the CHD dataset.

Table 3. The summary of the continuous variables in the CHD dataset.

Status of Coronary Heart Disease	Statistics	Age	Trestbps	Cholesterol	Thalach	Oldpeak
No	Min	29	94	126	96	0
	Q1	45	120	209	148.5	0
	Mean	52.67	129.20	243.06	158.29	0.59
	Mode	54	130	204	162	0
	Q3	59	140	267.5	172	1.05
	Max	76	180	564	202	4.2
Yes	Min	35	100	131	71	0
	Q1	52	120	217.5	125	0.55
	Mean	56.63	134.57	251.47	139.26	1.57
	Mode	58	140	254	132	0
	Q3	62	145	283.5	156.5	2.5
	Max	77	200	409	195	6.2

The seven categorical variables of the CHD dataset are depicted in Figure 4. For the class Yes, the distribution of data that is more common in each variable is male (sex), asymptomatic/fourth type (CP), fasting blood sugar less than 120 mg/dL (FBS), left ventricular hypertrophy (Restecg), angina due to exercise (Exang), a flat top of the ST segment (slope), and a reversible defect (Thal).

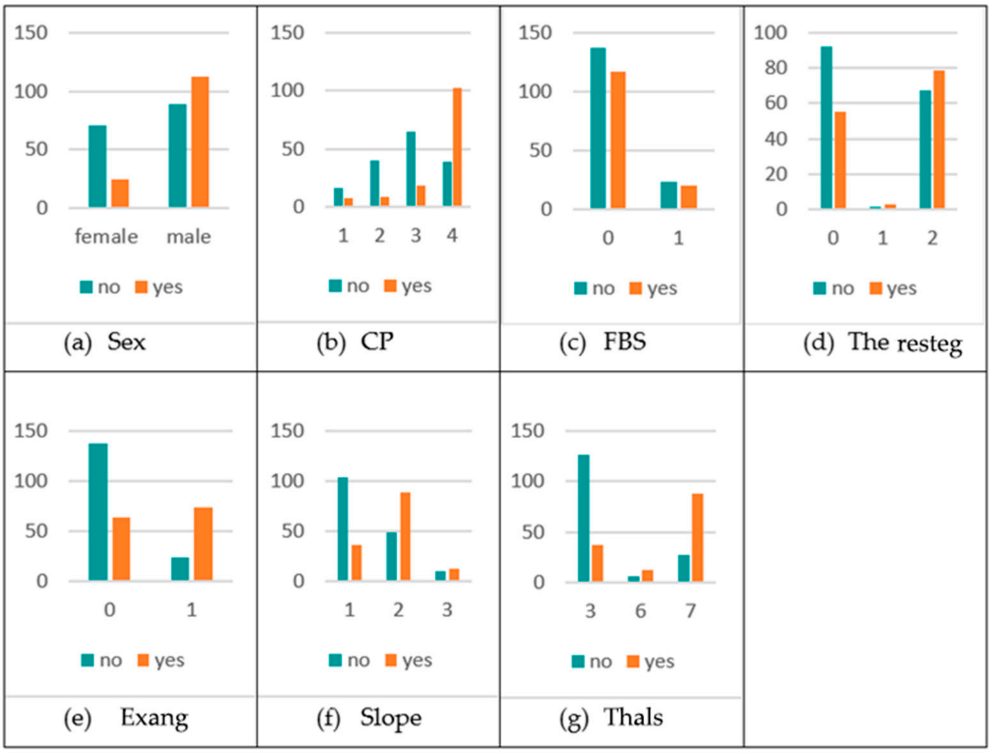


Figure 4. The summary of categorical variables for the CHD dataset.

3.1.2. Discretization

Crisp discretization for six numeric predictor variables in the CHD dataset is formed based on expert justification/prior information from the sources, as presented in Table 4 [33].

Table 4. Crisp Discretization based on Prior Information of CHD.

Variable	Crisp Discretization	Source of Prior Information
Age	<40 years 40–64 years ≥65 years	Woodward et al., 2012 in [33]
Trestbts	90–119 mmHg (Normal) 120–139 mmHg (Prehypertension) ≥140 mmHg (Hypertension)	Borghi et al., 2003 in [33]
Chol	<200 (Normal) 200–239 (High Limit); ≥240 (High)	Third Report of the National Cholesterol Education Program (NCEP), 2001 in [33]
Thalach	≤100 (Normal) >100 (Takikardi)	Palatini, 1999 in [33]
Oldpeak	<3.2 (No/Normal) ≥3.2 (Yes/Risk)	Riani et al., 2019 in [33]

The fuzzy discretization can be formed based crisp discretization as shown in Table 4. The use of a combination of fuzzy membership functions with the same pattern demonstrates symmetry. The discretization process divides each variable into three categories, except for the Thalach and Oldpeak variables, which are divided into two categories. For each variable in each combination, the first category employs a decreasing pattern, including linear and nonlinear functions, such as the sigmoid function. The second category employs fuzzy memberships with symmetrical curve shapes, such as triangular, trapezoidal, pi, and beta memberships. We assume the width of the domain interval from the curve's center point to its end to be the same size. Then, in the third category, discretization uses fuzzy membership with increasing patterns, both linear and nonlinear functions (sigmoid functions). Discretization results using the six combinations of fuzzy membership functions for coronary heart data can be seen in Table 5.

Table 5. Fuzzy Discretization Interval for the CHD Dataset.

Continuous Variable	Discretization Term	Discretization Interval					
		FDTID3-1	FDTID3-2	FDT D3-3	FDTID3-4	FDTID3-5	FDTID3-6
Age	Young	[29, 41]	[29, 42]	[29, 45]	[29, 55]	[29, 41]	[29, 41]
	Middle	[39, 65]	[38, 66]	[38, 66]	[45, 69]	[40, 64]	[41, 63]
	Old	[63, 77]	[62, 77]	[63, 77]	[61, 77]	[63, 77]	[63, 77]
Trestbps	Normal	[90, 121]	[90, 122]	[90, 124]	[90, 130]	[90, 120]	[90, 120]
	Pre-Hypertension	[119, 141]	[119, 142]	[120, 144]	[110, 150]	[119, 149]	[120, 138]
	Hypertension	[139, 200]	[139, 200]	[140, 200]	[130, 200]	[138, 200]	[138, 200]
Cholesterol	Normal	[126, 202]	[126, 202]	[126, 210]	[126, 220]	[126, 201]	[126, 200]
	High limit	[198, 242]	[200, 242]	[200, 252]	[200, 260]	[200, 240]	[200, 240]
	High	[238, 564]	[240, 54]	[240, 54]	[240, 54]	[239, 564]	[240, 564]
Thalach	Normal	[71, 102]	[71, 105]	[71, 121]	[71, 203]	[71, 101]	[71, 101]
	Taki Karbi	[98, 202]	[100, 202]	[100, 202]	[71, 203]	[99, 202]	[100, 202]
Oldpeak	No	[0, 4]	[0, 4]	[0, 6]	[0, 6]	[0, 4]	[0, 4]
	Yes	[2, 6]	[2, 6]	[0, 6]	[0, 6]	[2, 6]	[2, 6]

The six fuzzy membership function combination models FDTID3-1–FDTID3-6 in Table 5 provide different discretization parameters for each membership function. In the Age variable, the three models, FDTID3-1, FDTID3-5, and FDTID3-6, have the same parameters for the Young category. The FDTID3-2 and FDTID3-3 models have the same parameters for the middle category. For the old category, only the FDTID3-2 and FDTID3-4 models have parameters different from those of the other models. In the Trestbps variable, Models FDTID3-1 and FDTID3-2 have the same parameters for all categories except for the upper limit of the pre-hypertension interval. The FDTID3-5 and FDTID3-6 models have the same parameters, only in the hypertension category. In the Cholesterol variable, the three models (FDTID3-2, FDTID3-3, and FDTID3-4) have the same parameters for the high category. The two models, FDTID3-5 and FDTID3-6, have the same parameters for high limit and Normal categories. The four models in the Oldpeak variable, FDTID3-1, FDTID3-2, FDTID3-5, and FDTID3-6, have the same parameters for each category. The other two models have parameters in both categories.

3.1.3. Five-Fold Cross Validation

The data division into five folds resulted in the calculation into five iterations. The data composition in each iteration for 5-fold cross-validation is presented in Table 6.

Table 6. Data Composition in Each Fold of CHD.

Data	Status	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Learning	No	128	128	125	127	131
	Yes	110	109	113	110	107
	Sum	238	237	238	237	238
Testing	No	32	32	35	33	29
	Yes	27	28	24	27	30
	Sum	59	60	59	60	59
	Total	297	297	297	297	297

The training data at each iteration cover about 80% of the data and the rest as test data. Sampling using the principle of no replacement causes the test data to be unseen.

3.1.4. FDTID3 Modeling

Figure 5 presents the FDTID3-4 model for the first iteration in predicting CHD status. In this CHD data modeling, the decision is 0, meaning that the observation does not have CHD status, and the result is 1, meaning that the observation has CHD status. The figure shows that in the FDTID3-4 model, CP (type of chest pain) is the variable that most influences CHD status. Its position as the Root Node indicates this. The two variables in the first node are Thals and Ca. In the next node are Cholesterol, Resteg, Slope, Ca, Thal, Sex, and Trestbps. The earlier the node position, the greater the influence of variables at that node on the prediction of CHD status. In this first iteration of FDTID3-4, 81 rules for predicting CHD status exist. Each rule begins with the notation of $[R_w]$, $w = 1, 2, \cdots, 81$ and is presented below:

$[R_1]$ If CP is Typical Angina and Thal is Normal, then the decision is 0.

$[R_2]$ If CP is Typical Angina and Thal is Permanent disability, then the decision is 0.

$[R_3]$ If CP is Typical Angina, Thal is Temporary disability, and Cholesterol is a High Limit, then the decision is 1.

$[R_4]$ If CP is Typical Angina, Thal is Temporary disability, and Cholesterol is a High Limit, then the decision is 1.

$[R_5]$ If CP is Typical Angina, Thal is Temporary disability, Cholesterol is a High Limit, and the Trestbps is Normal, then the decision is 1.

$[R_6]$ If CP is Typical Angina, Thal is Temporary disability, Cholesterol is a High Limit, Trestbps is Hypertension, FBS is False, and Oldpeak is No, then the decision is 1.

$[R_7]$ If CP is Typical Angina, Thal is Temporary disability, Cholesterol is a High Limit, Trestbps is Hypertension, FBS is False, and Oldpeak is Yes, then the decision is 1.

$[R_8]$ If CP is typical angina, Thal is Temporary disability, Cholesterol is a High Limit, Trestbps is Hypertension, and FBS is True, then the decision is 0.

$[R_9]$ If CP is Atypical Angina, Thal is Normal, and Restecg is Normal, then the decision is 0.

$[R_{10}]$ If CP is Atypical Angina, Thal is Normal, Restecg is Ventricular hypertrophy, and Ca is 0, then the decision is 0.

$[R_{11}]$ If CP is Atypical Angina, Thal is Normal, Restecg is Ventricular hypertrophy, Ca is 1, and Age is Middle, then the decision is 1.

$[R_{12}]$ If CP is Atypical Angina, Thal is Normal, Restecg is Ventricular hypertrophy, Ca is 1, and Age is Old, then the decision is 0.

$[R_{13}]$ If CP is Atypical Angina, Thal is Normal, Restecg is Ventricular hypertrophy, and Ca is 1, then the decision is 1.

$[R_{14}]$ If CP is Atypical Angina and Thal is Permanent disability, then the decision is 1.

$[R_{15}]$ If CP is Atypical Angina, Thal is Temporary disability, Slope is Learning Up, and Trestbps is Hypertension, then the decision is 1.

$[R_{16}]$ If CP is Atypical Angina, Thal is Temporary disability, Slope is Learning Up, and Trestbps is Normal, then the decision is 0.

$[R_{17}]$ If CP has Atypical Angina, Thal has a Temporary disability, Slope is Learning Up, Trestbps is prehypertension, and Cholesterol is High, then the decision is 1.

$[R_{18}]$ If CP has Atypical Angina, Thal has a Temporary disability, Slope is Learning Up, Trestbps is prehypertension, and Cholesterol is High limit, then the decision is 0.

$[R_{19}]$ If CP is Atypical Angina, Thal is Temporary disability, and Slope is Flat, then the decision is 1.

$[R_{20}]$ If CP has Atypical Angina, Thal has Temporary disability, and the Slope is Slightly Sloping, then the decision is 1.

$[R_{21}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Hypertension, Oldpeak is No, and Age is Middle, then the decision is 0.

$[R_{22}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Hypertension, Oldpeak is No, Age is Old, and Sex is Male, then the decision is 0

$[R_{23}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Hypertension, Oldpeak is No, Age is Middle, and Sex is Female, then the decision is 1.

$[R_{24}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Hypertension, Oldpeak is No, and Age is Young, then the decision is 0.

$[R_{25}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Hypertension, and Oldpeak is Yes, then the decision is 1.

$[R_{26}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, and Sex is Male, then the decision is 0.

$[R_{27}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, Sex is Female, Cholesterol is High, Age is Middle, and Restecg is Normal, then the decision is 1.

$[R_{28}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, Sex is Female, Cholesterol is High, Age is Middle, and Restecg is Ventricular hypertrophy, then the decision is 0.

$[R_{29}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, Sex is Female, Cholesterol is High, and Age is Young, then the decision is 0.

$[R_{30}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, Sex is Female, and Cholesterol is High Limit, then the decision is 0.

$[R_{31}]$ If CP is Nonanginal pain, Thal is Normal, Ca is 0, Trestbps is Normal, Sex is Female, and Cholesterol is Normal, then the decision is 0.

$[R_{32}]$ If CP is Nonanginal pain, Thal is Normal, and Ca is 1, then the decision is 0.

$[R_{33}]$ If CP is Nonanginal pain, Thal is Normal, and Ca is 2, then the decision is 0.

$[R_{34}]$ If CP is Nonanginal pain, Thal is Normal, and Ca is 3, then the decision is 1.

$[R_{35}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Leaning up, and Trestbps is Hypertension, then the decision is 1.

$[R_{36}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Leaning up, and Trestbps is Hypertension, then the decision is 0.

$[R_{37}]$ If CP is Nonanginal pain and Thal is Temporary disability, then the decision is 1.

$[R_{38}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Leaning up, and Trestbps is Prehypertension, then the decision is 0.

$[R_{39}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, and Trestbps is Hypertension, then the decision is 1.

$[R_{40}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, Trestbps is Normal, and Ca is 0, then the decision is 0.

$[R_{41}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, Trestbps is Normal, and Ca is 1, then the decision is 1.

$[R_{42}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, Trestbps is Normal, and Ca is 3, then the decision is 1.

$[R_{43}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, Trestbps is Prehypertension, and Ca is 1, then the decision is 1.

$[R_{44}]$ If CP is Nonanginal pain, Thal is Temporary disability, Slope is Flat, Trestbps is Prehypertension, and Ca is 3, then the decision is 0.

$[R_{45}]$ If CP is Nonanginal pain, Thal is Temporary disability, and the Slope is Slightly Sloping, then the decision is 0.

$[R_{46}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, Sex is Male, Age is Middle, and Restecg is Normal, then the decision is 1.

$[R_{47}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, Sex is Male, Age is Middle, and Restecg is ST-T wave abnormalities, then the decision is 1.

$[R_{48}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, Sex is Male, Age is Middle, Restecg is Ventricular hypertrophy, and Slope is Leaning up, then the decision is 1.

$[R_{49}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, Sex is Male, Age is Middle, Restecg is Ventricular hypertrophy, and Slope is Flat, then the decision is 0.

$[R_{50}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, Sex is Male, and Age is Old, then the decision is 0.

$[R_{51}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Hypertension, and Sex is Female, then the decision is 0.

$[R_{52}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, and Trestbps is Normal, then the decision is 0.

$[R_{53}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Prehypertension, Age is Middle, and Restecg is Normal, then the decision is 0.

$[R_{54}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Prehypertension, Age is Middle, Restecg is Ventricular Hypertrophy, and Slope is Learning Up, then the decision is 1.

$[R_{55}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Prehypertension, Age is Middle, Restecg is Ventricular Hypertrophy, and Slope is Flat, then the decision is 1.

$[R_{56}]$ If CP is Asymptomatic, Ca is 0, Thal is Normal, Trestbps is Prehypertension, and Age is Young, then the decision is 1.

$[R_{57}]$ If CP is Asymptomatic, Ca is 0, Thal is Permanent Disability, and Trestbps is Hypertension, then the decision is 0.

$[R_{58}]$ If CP is Asymptomatic, Ca is 0, Thal is Permanent Disability, Trestbps is Normal, and Cholesterol is High Limit, then the decision is 0.

$[R_{59}]$ If CP is Asymptomatic, Ca is 0, Thal is Permanent Disability, Trestbps is Normal, and Cholesterol is Normal, then the decision is 1.

$[R_{60}]$ If CP is Asymptomatic, Ca is 0, Thal is Permanent Disability, and Trestbps is Prehypertension, then the decision is 1.

$[R_{61}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, and Cholesterol is High, then the decision is 1.

$[R_{62}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is High Limit, Age is Middle, and Fbs is False, then the decision is 0.

$[R_{63}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is High Limit, Age is Middle, and Fbs is True, then the decision is 1.

$[R_{64}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is High Limit, and Age is Young, then the decision is 1.

$[R_{65}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is Normal, and Age is Middle, then the decision is 1.

$[R_{66}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is High Limit, Age is Old, and Trestbps is Hypertension, then the decision is 1.

$[R_{67}]$ If CP is Asymptomatic, Ca is 0, Thal is Temporary Disability, Cholesterol is High Limit, Age is Old, and Trestbps is Normal, then the decision is 0.

$[R_{68}]$ If CP is Asymptomatic, Ca is 0, Thal is a Temporary Disability, Cholesterol is Normal, and Age is Young, then the decision is 1.

$[R_{69}]$ If CP is Asymptomatic, Ca is 1, and Sex is Male, then the decision is 0.

$[R_{70}]$ If CP is Asymptomatic, Ca is 1, Sex is Female, and Age is Middle, then the decision is 1.

$[R_{71}]$ If CP is Asymptomatic, Ca is 1, Sex is Female, and Age is Old, then the decision is 0.

$[R_{72}]$ If CP is Asymptomatic, Ca is 1, Sex is Male, Age is Old, and Restecg is Ventricular Hypertrophy, then the decision is 1.

$[R_{73}]$ If CP is Asymptomatic and Ca is 1, and Sex is Male, and Age is Young, then the decision is 1.

$[R_{74}]$ If CP is Asymptomatic, Ca is 2, Thal is Normal, and Cholesterol is High, then the decision is 1.

$[R_{75}]$ If CP is Asymptomatic, Ca is 2, Thal is Normal, and Cholesterol is High Limit, then the decision is 0.

$[R_{76}]$ If CP is Asymptomatic, Ca is 2, and Thal is Permanent disability, then the decision is 1.

$[R_{77}]$ If CP is Asymptomatic, Ca is 2, and Thal is Temporary disability, then the decision is 1.

$[R_{78}]$ If CP is Asymptomatic, Ca is 3, and Trestbps is Hypertension, then the decision is 1.

$[R_{79}]$ If CP is Asymptomatic, Ca is 3, Trestbps is Normal, and Cholesterol is High, then the decision is 1.

$[R_{80}]$ If CP is Asymptomatic, Ca is 3, Trestbps is Normal, and Cholesterol is High Limit, then the decision is 0.

$[R_{81}]$ If CP is Asymptomatic, Ca is 3, and Trestbps is Prehypertension, then the decision is 1.

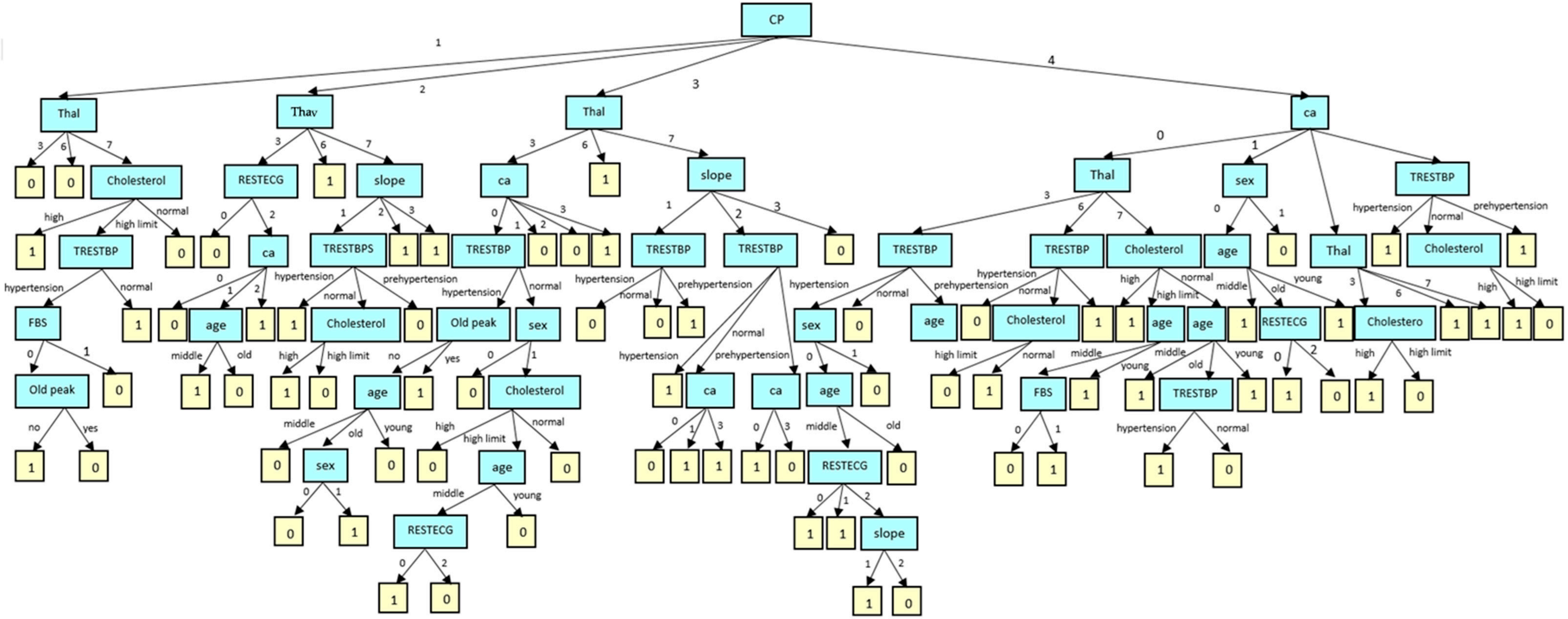


Figure 5. The first iteration of the FDTID3-4 model with the complete variable for CHD status prediction. The confusion matrix for the first iteration of FDTID3-4 with 13 predictor variables (complete) is presented in Table 7.

Table 7. Confusion matrix of CHD status prediction of the first iterations of FDTID3-4 with 13 predictor variables.

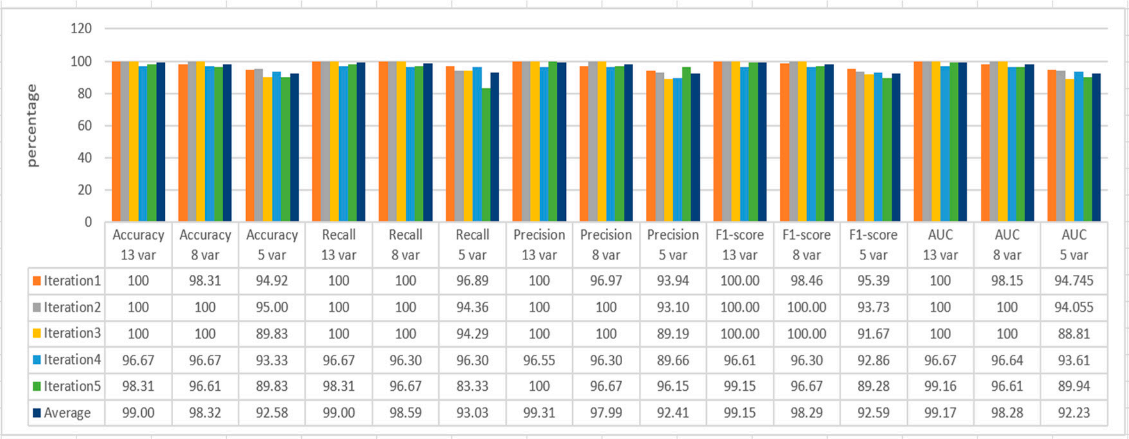
		Prediction of CHD Status		Sum
The fact of CHD Status	Yes	27	0	27
	No	0	32	32
	Sum	27	32	59

The first and third rules are examples of decisions that make predictions of 0 and 1, respectively. In the first rule, if someone is in a condition of Typical Angina chest pain and has no blood disorders (Thal), their status is predicted not to have CHD. In the third rule, if someone is in a condition of Typical Angina chest pain, has thalassemia type temporary disability, and cholesterol in the High Limit category, then it is predicted that their status is CHD.

The confusion matrix in Table 7 shows that True Positive prediction (TP) is 27, True Negative prediction (TN) is 32, and False Positive (FP) and False Negative (FN) predictions are each 0. Because FP and FN are all 0, or the predictions are all correct for both class 0 and class 1, then all metrics are worth 100.

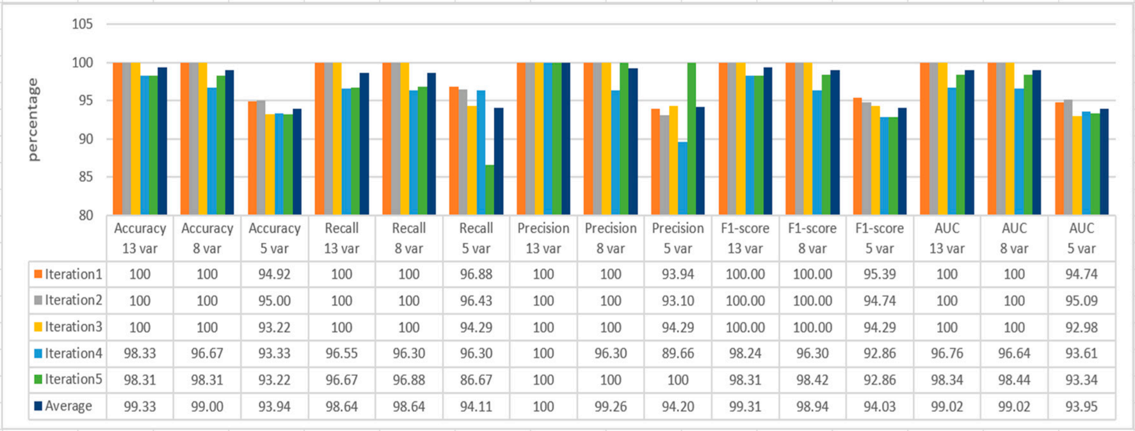
In the other four iterations, the significant variables at the root node to the second node consist of eight variables: CP, Thal, Ca, Cholesterol, Resteg, Slope, Sex, and Trestbps. However, the last five variables are not found at the root and first nodes. This pattern is also found in the FDTID3 model with five other fuzzy membership function combinations. For this reason, in this research, the FDTID3-1 to FDTID3-6 models for predicting CHD status were built using complete predictor variables (13 variables) and 8 variables and 5 variables as the result of variable selection based on Node position. The evaluation of the prediction models FDTID3-1 to FDTID3-6, each based on the number of predictor variables, is presented in Figures 6–11.

In the FDTID3-1 model, each iteration shows the value of all metrics above 85%. Several metrics in several iterations have a value of 100. Even in the first and third iterations, all metrics have a value of 100. Accuracy, recall, and AUC have the same value range of 96.67–100%, while precision and F1-score each have a value range of 96.55–100% and 99.16–100%. The average value of each metric from the five iterations in the FDTID3-1 model with 13 predictor variables is the highest compared to the other two FDTID3-1 models, which involve 8 and f5 variables, respectively. Next, the other five fuzzy combination models, namely FDTID3-2–FDTID3-6, were each given the same treatment as the FDTID3-1 model, where the model metric measures were explored based on complete variables and variable selection involving only 8 and 5 variables, respectively.



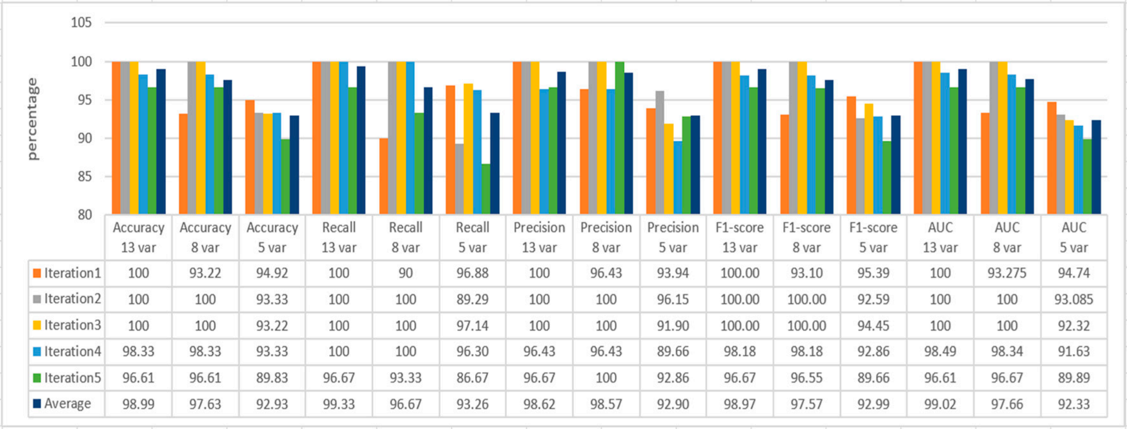
	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
Iteration1	100	98.31	94.92	100	100	96.89	100	96.97	93.94	100.00	98.46	95.39	100	98.15	94.745
Iteration2	100	100	95.00	100	100	94.36	100	100	93.10	100.00	100.00	93.73	100	100	94.055
Iteration3	100	100	89.83	100	100	94.29	100	100	89.19	100.00	100.00	91.67	100	100	88.81
Iteration4	96.67	96.67	93.33	96.67	96.30	96.30	96.55	96.30	89.66	96.61	96.30	92.86	96.67	96.64	93.61
Iteration5	98.31	96.61	89.83	98.31	96.67	83.33	100	96.67	96.15	99.15	96.67	89.28	99.16	96.61	89.94
Average	99.00	98.32	92.58	99.00	98.59	93.03	99.31	97.99	92.41	99.15	98.29	92.59	99.17	98.28	92.23

Figure 6. Performance of FDTID3-1 of CHD based on 5-fold cross-validation.



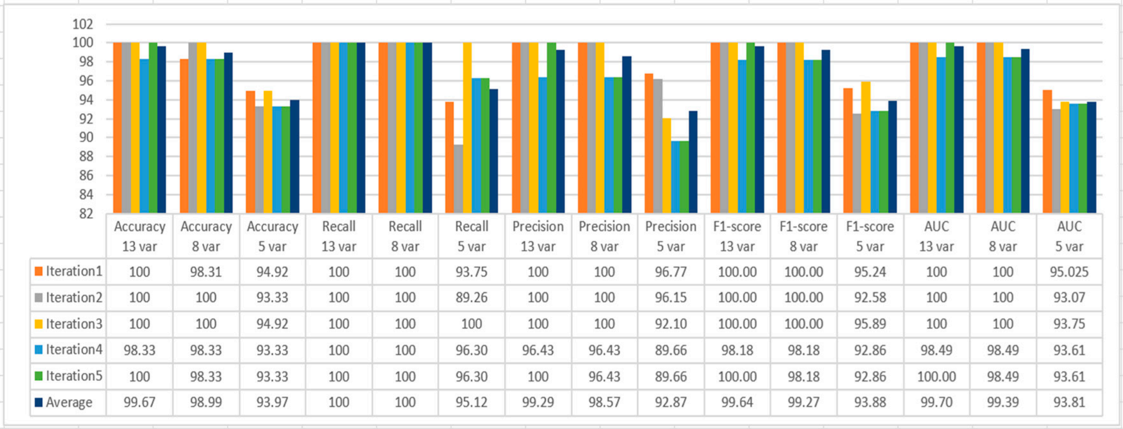
	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
■ Iteration1	100	100	94.92	100	100	96.88	100	100	93.94	100.00	100.00	95.39	100	100	94.74
■ Iteration2	100	100	95.00	100	100	96.43	100	100	93.10	100.00	100.00	94.74	100	100	95.09
■ Iteration3	100	100	93.22	100	100	94.29	100	100	94.29	100.00	100.00	94.29	100	100	92.98
■ Iteration4	98.33	96.67	93.33	96.55	96.30	96.30	100	96.30	89.66	98.24	96.30	92.86	96.76	96.64	93.61
■ Iteration5	98.31	98.31	93.22	96.67	96.88	86.67	100	100	100	98.31	98.42	92.86	98.34	98.44	93.34
■ Average	99.33	99.00	93.94	98.64	98.64	94.11	100	99.26	94.20	99.31	98.94	94.03	99.02	99.02	93.95

Figure 7. Performance of FDTID3-2 of CHD based on 5-fold cross-validation.



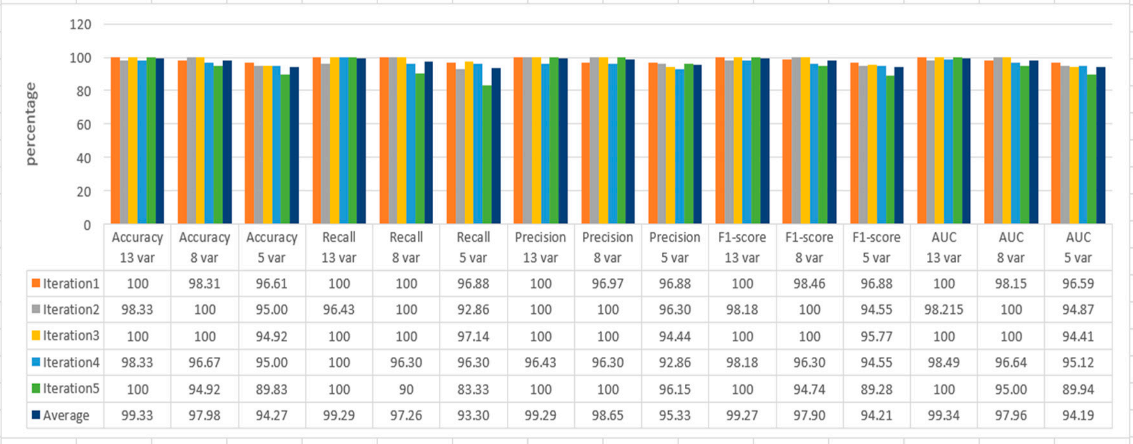
	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
■ Iteration1	100	93.22	94.92	100	90	96.88	100	96.43	93.94	100.00	93.10	95.39	100	93.275	94.74
■ Iteration2	100	100	93.33	100	100	89.29	100	100	96.15	100.00	100.00	92.59	100	100	93.085
■ Iteration3	100	100	93.22	100	100	97.14	100	100	91.90	100.00	100.00	94.45	100	100	92.32
■ Iteration4	98.33	98.33	93.33	100	100	96.30	96.43	96.43	89.66	98.18	98.18	92.86	98.49	98.34	91.63
■ Iteration5	96.61	96.61	89.83	96.67	93.33	86.67	96.67	100	92.86	96.67	96.55	89.66	96.61	96.67	89.89
■ Average	98.99	97.63	92.93	99.33	96.67	93.26	98.62	98.57	92.90	98.97	97.57	92.99	99.02	97.66	92.33

Figure 8. Performance of FDTID3-3 of CHD based on 5-fold cross-validation.



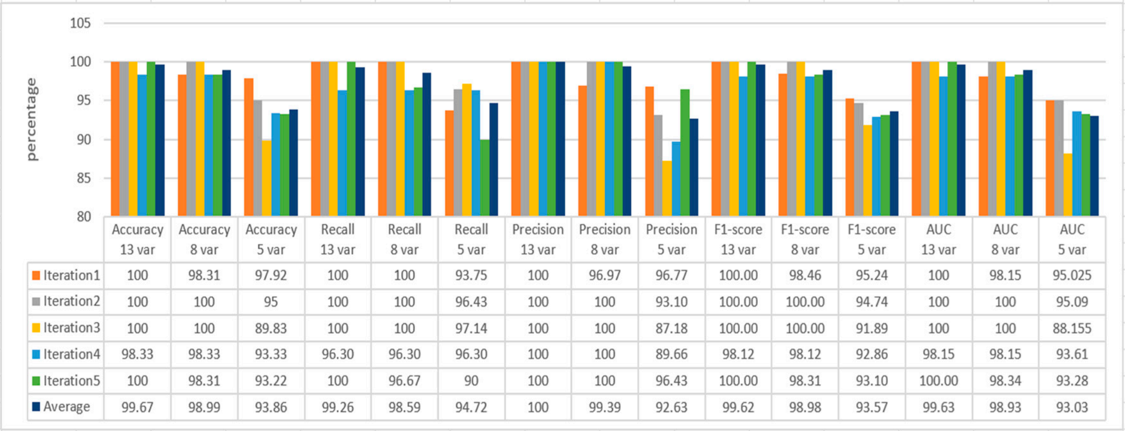
	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
■ Iteration1	100	98.31	94.92	100	100	93.75	100	100	96.77	100.00	100.00	95.24	100	100	95.025
■ Iteration2	100	100	93.33	100	100	89.26	100	100	96.15	100.00	100.00	92.58	100	100	93.07
■ Iteration3	100	100	94.92	100	100	100	100	100	92.10	100.00	100.00	95.89	100	100	93.75
■ Iteration4	98.33	98.33	93.33	100	100	96.30	96.43	96.43	89.66	98.18	98.18	92.86	98.49	98.49	93.61
■ Iteration5	100	98.33	93.33	100	100	96.30	100	96.43	89.66	100.00	98.18	92.86	100.00	98.49	93.61
■ Average	99.67	98.99	93.97	100	100	95.12	99.29	98.57	92.87	99.64	99.27	93.88	99.70	99.39	93.81

Figure 9. Performance of FDTID3-4 of CHD based on 5-fold cross-validation.



	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
■ Iteration1	100	98.31	96.61	100	100	96.88	100	96.97	96.88	100	98.46	96.88	100	98.15	96.59
■ Iteration2	98.33	100	95.00	96.43	100	92.86	100	100	96.30	98.18	100	94.55	98.215	100	94.87
■ Iteration3	100	100	94.92	100	100	97.14	100	100	94.44	100	100	95.77	100	100	94.41
■ Iteration4	98.33	96.67	95.00	100	96.30	96.30	96.43	96.30	92.86	98.18	96.30	94.55	98.49	96.64	95.12
■ Iteration5	100	94.92	89.83	100	90	83.33	100	100	96.15	100	94.74	89.28	100	95.00	89.94
■ Average	99.33	97.98	94.27	99.29	97.26	93.30	99.29	98.65	95.33	99.27	97.90	94.21	99.34	97.96	94.19

Figure 10. Performance of FDTID3-5 of CHD based on 5-fold cross-validation.



	Accuracy 13 var	Accuracy 8 var	Accuracy 5 var	Recall 13 var	Recall 8 var	Recall 5 var	Precision 13 var	Precision 8 var	Precision 5 var	F1-score 13 var	F1-score 8 var	F1-score 5 var	AUC 13 var	AUC 8 var	AUC 5 var
■ Iteration1	100	98.31	97.92	100	100	93.75	100	96.97	96.77	100.00	98.46	95.24	100	98.15	95.025
■ Iteration2	100	100	95	100	100	96.43	100	100	93.10	100.00	100.00	94.74	100	100	95.09
■ Iteration3	100	100	89.83	100	100	97.14	100	100	87.18	100.00	100.00	91.89	100	100	88.155
■ Iteration4	98.33	98.33	93.33	96.30	96.30	96.30	100	100	89.66	98.12	98.12	92.86	98.15	98.15	93.61
■ Iteration5	100	98.31	93.22	100	96.67	90	100	100	96.43	100.00	98.31	93.10	100.00	98.34	93.28
■ Average	99.67	98.99	93.86	99.26	98.59	94.72	100	99.39	92.63	99.62	98.98	93.57	99.63	98.93	93.03

Figure 11. Performance of FDTID3-6 of CHD based on 5-fold cross-validation.

The performance of the FDTID3-2 model, as presented in Figure 7, shows that most metrics involving all variables in the prediction process have the highest values compared to the other two FDTID3-2 models. In this model, all metrics give a value of 100 from the first to third iterations. The average of the five iterations shows that the FDTID3-2 model with 13 variables has the highest accuracy, precision, and F1 score.

As with the FDTID3-1 model, the performance of the FDTID3-3 model, as shown in Figure 8, shows that the model involving all variables in the prediction process has the highest value on all metrics, even having a value of 100 from the first to third iterations as in the FDTID3-2 model. The fact is followed successively by models involving eight and five variables as the results of predictor variable selection.

Similar events occur in the FDTID3-4 model, as shown in Figure 9, where the FDTID3-4 model involving 13 predictor variables has the highest average performance compared to the other two FDTID3-4 models. Although the recall in the FDTID3-4 model with eight variables also has a value of 100, the other four metrics are not higher than the FDTID3-4 model with thirteen variables.

In the FDTID3-5 model, the five-evaluation metrics (Figure 10) show similar events to the FDTID3-1–FDTID3-4 models, where all average metric values indicate that the FDTID3-5 model with 13 predictor variables is the best compared to the other two FDTID3 models. In this model, all metrics also give a value of 100 from the first to third iterations.

Like the FDTID3-1–FDTID3-5 model, in the FDTID3-6 model, the performance of the model with complete predictor variables is the model that predicts CHD status with the best performance. All metrics in the model have metrics with perfect values from iteration 1 to iteration 3 (Figure 11). The information indicates that the FDTID3 model with 13 variables is the best model for predicting CHD status. This study also compares the performance of the DTID3 model with 13 variables using the concept of crisp sets in discretizing predictor variables.

The comparison of the performance of the DTID3 and FDTID3-1–FDTID3-6 models involving all predictor variables in predicting CHD status is summarized in Table 8. The value of the metrics of each model is the average of the 5-fold performance.

Table 8. The prediction performance of CHD status.

Fuzzy Membership Functions Combination	Prediction Performance Metric (%)				
	Accuracy	Recall	Precision	F1-Score	AUC
DTID3	98.99	99.33	98.62	98.97	99.02
FDTID3-1	99.00	99.00	99.31	99.15	99.17
FDTID3-2	99.33	98.64	100	99.31	99.02
FDTID3-3	98.99	99.33	98.62	98.97	99.02
FDTID3-4	99.67	100	99.29	99.64	99.70
FDTID3-5	99.33	99.29	99.29	99.27	99.34
FDTID3-6	99.67	99.26	100	99.62	99.63

In predicting disease states, including CHD, a high recall value is often considered better than a high precision value. This is because it is assumed that a model is better when predicting a patient's status as positively sick. However, if the patient's status is healthy (negative), then a model that predicts a patient's status as healthy. However, the patient's status is positive. However, higher false positives are not always better than higher false negatives when predicting disease status. A person who is predicted to be positive when they are not may be detrimental to the patient because it can cause stress or other excessive responses. A higher F1 score is better because this metric balances False Positives and False Negatives. All the proposed FDTID3 models performed better than the DTID3 model except the FDTID3-3 model, which performed the same as the DTID3 model. The FDTID3-4 model had the highest three metric values: recall, F1-score, and AUC. The number recorded the FDTID3-4 model as the FDTID3 model with the highest metric value compared to other FDTID3 models. Therefore, the FDTID3-4 model with 13 predictor variables is the best model for predicting CHD status. This fact informs us that most of the first and third categories in each variable tend to have decreasing sigmoid and increasing sigmoid functions rather than decreasing linear and increasing linear. Thus, the second category tends to have a triangular function rather than a pi function if the other two categories are sigmoid functions.

Table 9 presents the analysis of variance (ANOVA) for the seven models (six models of FDTID3 and one model of DTID3). Whether the performance of the seven proposed models is distinct from that of the others, it is worthwhile to compare them using Monte Carlo resampling [35]. The ANOVA demonstrates that the seven proposed models differ in at least one average performance metric for accuracy, precision, recall, F1-score, and AUC at 5% significance levels.

Moreover, it examines which model pairs perform significantly differently and whether the DTID3 classification performance metrics have improved when using the proposed FDTID3. Post hoc tests with a significance level of 5% using Tukey–Kramer are given in Table 10.

Most model pairs have an absolute mean difference (AMD) that exceeds each metric's Q-critical value, namely 0.16, 0.26, 0.22, 0.19, and 0.2. Only the performance of the FDTID3-1 and DTID3 model pairs is not significantly different on all five metrics. In comparison, the other five FDTID3 models are significantly different from each other on at least two metrics. This fact also indicates that the five FDTID3 models have significantly different (increased) performance from DTID3. In addition, considering that the other five FDTID3 models have different fuzzy membership functions, it can be concluded that the performance of the proposed FDTID3 model depends on the fuzzy membership function used. Our hypothesis that the seven proposed models are different has been proven.

Table 9. ANOVA of the proposed model for the CHD dataset.

Metrics	Source of Var.	Sum of Squares	Mean Squares	F	*p*-Value	F-Criteria
Accuracy	between	256.06	42.68	645.90	6.0×10^{-186}	2.12
	within	23.13	0.07			
Recall	between	181.26	30.21	165.11	7.3×10^{-99}	
	within	64.04	0.18			
Precision	between	356.08	59.35	455.94	4.6×10^{-162}	
	within	45.56	0.13			
F1-score	between	208.24	34.71	351.64	7.3×10^{-145}	
	within	34.55	0.10			
AUC	between	248.89	41.48	387.31	3.4×10^{-151}	
	within	37.49	0.11			

Table 10. Tukey–Kramer test of the proposed model for the CHD dataset.

Comparison Model	Absolute Mean Difference				
	Accuracy	Recall	Precision	F1-Score	AUC
FDTID3-1 vs. FDTID3-2	1.47	1.12	1.60	1.36	1.50
FDTID3-1 vs. FDTID3-3	1.91	1.12	2.41	1.77	1.97
FDTID3-1 vs. FDTID3-4	1.47	1.93	0.82	1.37	1.43
FDTID3-1 vs. FDTID3-5	0.52	1.12	0.13	0.49	0.56
FDTID3-1 vs. FDTID3-6	2.25	1.83	2.15	1.99	2.17
FDTID3-1 vs. DTID3	0.00	0.02	0.00	0.01	0.01
FDTID3-2 vs. FDTID3-3	0.43	0.00	0.81	0.40	0.47
FDTID3-2 vs. FDTID3-4	0.00	0.81	0.78	0.01	0.07
FDTID3-2 vs. FDTID3-5	0.96	0.00	1.73	0.87	0.94
FDTID3-2 vs. FDTID3-6	0.78	0.71	0.55	0.63	0.67
FDTID3-2 vs. DTID3	1.47	1.10	1.60	1.36	1.49
FDTID3-3 vs. FDTID3-4	0.43	0.80	1.59	0.40	0.54
FDTID3-3 vs. FDTID3-5	1.39	0.00	2.54	1.28	1.41
FDTID3-3 vs. FDTID3-6	0.35	0.71	0.25	0.23	0.20
FDTID3-3 vs. DTID3	1.91	1.11	2.41	1.76	1.97
FDTID3-4 vs. FDTID3-5	0.96	0.81	0.95	0.88	0.87
FDTID3-4 vs. FDTID3-6	0.78	0.10	1.33	0.62	0.74
FDTID3-4 vs. DTID3	1.47	1.91	0.82	1.36	1.43
FDTID3-5 vs. FDTID3-6	1.74	0.71	2.28	1.50	1.61
FDTID3-5 vs. DTID3	0.52	1.10	0.13	0.48	0.56
FDTID3-6 vs. DTID3	1.74	0.71	2.28	1.50	1.61

3.2. Diabetes Mellitus Disease Dataset

3.2.1. Dataset Exploration and Preprocessing

In the DMD, the majority of predictor variables have zero data, and some studies assume that these zero data are missing data, so they focus on dealing with missing data first and then perform prediction/classification tasks and improve the performance of the previous model [27,28]. In addition, some researchers try to balance the classes distributed into classes with DM status (65%) and classes that do not have DM status (35%).

The dataset has eight quantitative predictor variables, seven continuous, and only one discrete. There are no missing data values in all variables, but the variables of Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, and Pregnancy have zero values. Except for Skin Thickness, Pregnancy, and BMI, zero values in these variables can be a critical and dangerous medical situation for the diabetic patient. Hence, some studies still leave them as zero in the prediction process [23,46]. However, some studies consider them missing data and impute them with the mean value [27,28]. In this study, the values in these variables are still left as zero because the DTID3 method can handle zero values.

The histogram of each continuous predictor variable and the bar plot of Pregnancies are given in Figure 12. None of the seven continuous predictor variables show a normal distribution. The majority of the distributions are skewed to the right. The bar plot of pregnancies shows that the number of women who have never been pregnant has a reasonably high frequency, but pregnancy of one is the highest frequency. The number continues to decrease as the frequency increases.

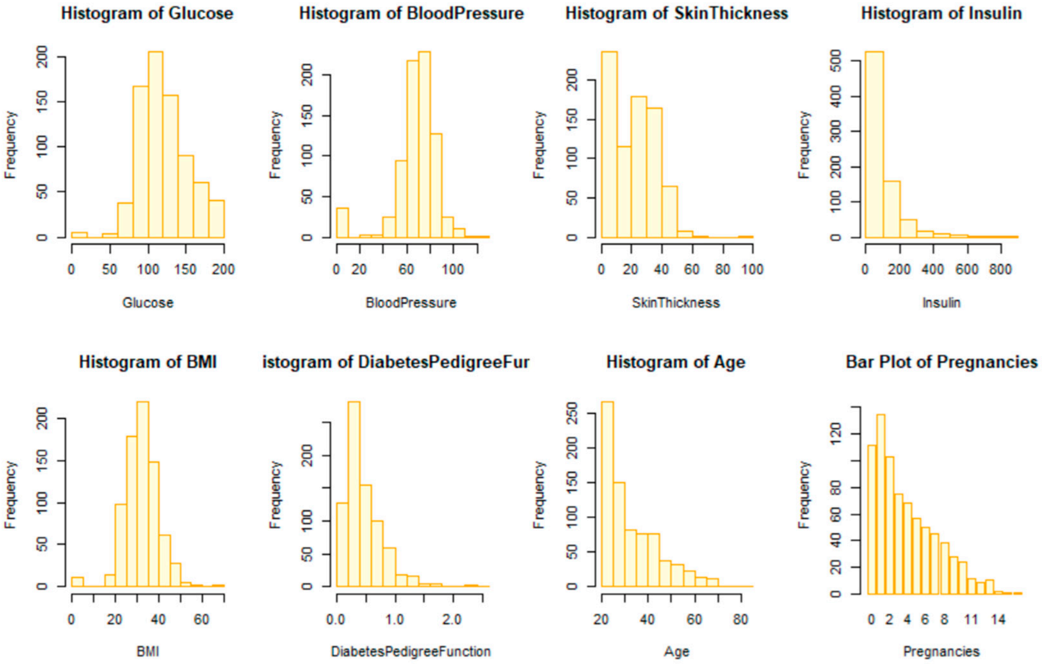


Figure 12. Histogram and Bar Plot of Predictor Variable in DM.

For each Yes and No class, a summary of the predictor variables of the DMD is presented in Table 11.

Table 11. The summary of continuous variables in the diabetes dataset.

Status of Diabetes	Stat.	Glucose (mg/dL)	Blood Pressure (mmHg)	Skin Thickness (mm)	Insulin (µ/mL)	BMI (kg/hg)	Diabetes Pedigree Function (unit)	Age (Year)	Pregnancies
No	Min	0	0	0	0	0	0.08	21	0
	Q1	93	62	0	0	25.4	0.23	23	1
	Mean	109.98	68.18	19.66	68.79	30.30	54.73	31.19	3.30
	Mode	99	74	0	0	0	0.207	22	1
	Q3	125	78	31	105	35.3	0.56	37	5
	Max	197	122	60	744	57.3	2329.00	81	13
Yes	Min	0	0	0	0	0	0.09	21	0
	Q1	119	66	0	0	30.8	0.26	28	1.75
	Mean	141.26	70.82	22.16	100.34	35.14	131.80	37.07	4.87
	Mode	125	70	0	0	32.9	0.254	25	0
	Q3	167	82	36	167.25	38.775	0.73	44	8
	Max	199	114	99	846	67.1	2288.00	70	17

The mean value of all predictor variables is higher in the class Yes. However, class No has the maximum values for variables such as Blood Pressure, Diabetes Pedigree Function, and Age. The zero value for the six variables other than Diabetes Degree Function and Age is owned by the Yes and No classes. In this work, the zero data are left as zero because the DTID3 method has no problem with zero data.

3.2.2. Discretization

The crisp discretization for eight numeric predictor variables in the DMD is formed based on prior information from the sources, as presented in Table 12.

Table 12. Crisp discretization based on prior information of DMD.

Variable	Crisp Discretization	Source of Prior Information
Glucose	$<$140 mg/dL $\geq$140 mg/dL	Araki et al., 2020 in [46]
Blood Pressure	60–80 mm hg 81–89 mm hg $\geq$90 mm hg	Tsujimoto and Kajio, 2018 in [46]
Skin Thickness	$\leq$30 mm $>$30 mm	Marrodan et al., 2015 and Khadilkar et al., 2015 in [46]
Insulin Level	1–283 µU/mL 284–565 µU/mL 566–846 µU/mL	Equation (1)
BMI	$<$30 kg/m^2 $\geq$30 kg/m^2	Nutall, 2015 in [46]
Diabetes Pedigree Function	<0.4 0.4–0.8 >0.8	Survey, 2017 in [46]
Age	$\leq$35 years $>$35 years	Lampinen et al., 2009 in [46]
Pregnancies	$\leq$4 times $>$4 times	Karegowda et al., 2012 in [46]

The crisp discretization divides each variable into three categories, except for the Thalach and Oldpeak variables, which are divided into two categories. For each variable in each combination, the first category employs a decreasing pattern, including linear and nonlinear functions, such as the sigmoid function. The second category employs fuzzy memberships with symmetrical curve shapes, such as triangular, trapezoidal, pi, and beta memberships. We assume the width of the domain interval from the curve's center point to its end to be the same size. Then, in the third category, discretization uses fuzzy membership with increasing patterns, both linear and nonlinear functions (sigmoid functions).

Table 13 displays the discretization outcomes achieved using a blend of fuzzy membership functions on the DMD. This study also proposes six combinations of fuzzy membership functions for discretization for the dataset, the same as the CHD dataset. As in the CHD dataset, the six fuzzy membership function combination models FDTID3-1–FDTID3-6 for the diabetes disease dataset, as presented in Table 5, provide different discretization parameters for each membership function. For example, in the Glucose variable, the two models, FDTID3-1 and FDTID3-2, have the same parameters for all categories. Likewise, the two models FDTID3-3 and FDTID3-4.

Meanwhile, the FDTID3-5 and FDTID3-6 models have parameters that tend to be different. In the Blood Pressure variable, for the Normal category, every one of the three pairs of models has the same parameters. For the Pre-Hypertension category, three models have different parameters, namely FDTID3-1, FDTID3-3, and FDTID3-4, while in the Hypertension category, four models have four different parameters, namely FDTID3-3, FDTID3-4, FDTID3-5, and FDTID3-6.

The data composition in each iteration for 5-fold cross-validation is presented in Table 14. Dividing the data into five folds results in the training data in each iteration covering around 80% of the data and the remainder as test data.

Table 13. Discretization Interval for DMD.

Continuous	Discretization	Discretization Interval					
Variable	**Term**	**FDTID3-1**	**FDTID3-2**	**FDTID3-3**	**FDTID3-4**	**FDTID3-5**	**FDTID3-6**
Glucose	Low	[44, 60]	[44, 60]	[44, 62]	[44, 62]	[44, 64]	[44, 66]
	Normal	[60, 140]	[60, 140]	[59, 141]	[59, 141]	[58, 142]	[57, 143]
	High	[140, 200]	[140, 200]	[138, 200]	[138, 200]	[136, 200]	[134, 200]
Blood Pressure	Normal	[24, 80]	[24, 80]	[24, 82]	[24, 82]	[24, 84]	[24, 84]
	Pre-Hypertension	[79, 91]	[80, 90]	[79, 91]	[79, 91]	[77, 93]	[77, 93]
	Hypertension	[90, 122]	[90, 122]	[88, 122]	[88, 122]	[86, 122]	[86, 122]
Skin Thickness	Normal	[7, 30]	[7, 31]	[7, 31]	[7, 33]	[7, 32]	[7, 35]
	Thick	[30, 99]	[29, 99]	[28, 100]	[28, 99]	[28, 99]	[24, 99]
Insulin Level	Normal	[0, 166]	[0, 167]	[0, 166]	[0, 168]	[0, 168]	[0, 170]
	High	[166, 846]	[159, 846]	[166, 846]	[164, 846]	[164, 846]	[162, 846]
BMI	Normal	[18, 30]	[18, 31]	[18, 30]	[18, 32]	[18, 32]	[18, 36]
	Obesity	[30, 68]	[29, 68]	[30, 68]	[28, 68]	[28, 68]	[24, 68]
Diabetes Pedigree Function	Low	[0, 0.4]	[0, 0.4]	[0, 0.4]	[0, 0.4]	[0, 0.5]]	[0, 0.6]
	Normal	[0.4, 0.8]	[0.4, 0.8]	[0.4, 0.8]	[0.4, 0.8]	[0.2, 1]	[0.3, 0.9]
	High	[0.8, 2329]	[0.8, 2329]	[0.8, 2329]	[0.8, 2329]	[0.7, 2329]	[0.6, 2329]
Age	Young	[21, 35]	[21, 36]	[21, 35]	[21, 37]	[21, 39]	[21, 39]
	Old	[35, 81]	[34, 81]	[35, 81]	[33, 81]	[31, 81]	[31, 81]
Pregnancies	Normal	[1, 4]	[1, 5]	[1, 5]	[1, 5]	[1, 6]	[1, 7]
	High	[4, 17]	[3, 17]	[3, 17]	[3, 17]	[2, 17]	[3, 17]

Table 14. Data composition in each fold of DMD.

Data	Status	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Learning	No	400	400	397	401	402
	Yes	214	214	218	213	213
	Sum	614	614	615	614	615
Testing	No	100	100	103	99	98
	Yes	54	54	50	55	55
	Sum	154	154	153	154	153
	Total	768	768	768	768	768

Figure 13 presents the FDTID3-5 model for the first iteration of five-fold cross-validation predicting DM status. In this DM data modeling, the decision is 0, meaning that the observation does not have DM status, and the result is 1, meaning that the observation has DM status. The figure shows that in the FDTID3-5 model, BMI is the variable that most influences DM status. The three variables in the first node are Glucose, Skin Thickness, and Blood Pressure. The second nodes are pregnancy, age, blood pressure, and glucose. The earlier the node position, the greater the influence of variables at that node on CHD status prediction. In this first iteration of FDTID3-4, there are 110 rules for predicting CHD status. Each rule begins with the notation of $[R_w]$, $w = 1,\ 2, \cdots,\ 110$ and is presented below:

$[R_1]$ If BMI is Normal, Glucose is High, Pregnancy is High, and Skin Thickness is Normal, then the decision is 0.

$[R_2]$ If BMI is Normal, Glucose is High, Pregnancy is High, and Skin Thickness is Thick, then the decision is 1.

$[R_3]$ If BMI is Normal, Glucose is High, Pregnancy is Low, and Skin Thickness is Normal, then the decision is 0.

$[R_4]$ If BMI is Normal, Glucose is High, Pregnancy is Low, and Skin Thickness is Thin, then the decision is 1.

$[R_5]$ If BMI is Normal, Glucose is High, Pregnancy is Low, and Skin Thickness is Very Thin, then the decision is 0.

$[R_6]$ If BMI is Normal, Glucose is High, and Pregnancy is Normal, then the decision is 1.

$[R_7]$ If BMI is Normal, Glucose is Low, and Pregnancy is Low, then the decision is 0.

$[R_8]$ If BMI is Normal, Glucose is Low, Pregnancy is Normal, and Skin Thickness is Thick, then the decision is 1.

$[R_9]$ If BMI is Normal, Glucose is Low, Pregnancy is Normal, and Skin Thickness is Very Thin, then the decision is 0.

$[R_{10}]$ If BMI is Normal and Glucose is Low, then the decision is 0.

$[R_{11}]$ If BMI is Obesity, Glucose is High, and Age is Middle, then the decision is 1.

$[R_{12}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Normal, and Blood Pressure is Low, then the decision is 1.

$[R_{13}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Normal, and Blood Pressure is Prehypertension, then the decision is 0.

$[R_{14}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Thick, and Blood Pressure is Hypertension, then the decision is 0.

$[R_{15}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Thick, Blood Pressure is Prehypertension, and Pregnancy is High, then the decision is 0.

$[R_{16}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Thick, and Blood Pressure is Prehypertension, Pregnancies are Low, then the decision is 0.

$[R_{17}]$ If BMI is Obesity, Glucose is High, Age is Young, Skin Thickness is Thick, Blood Pressure is Prehypertension, and Pregnancy is Normal, then the decision is 1.

$[R_{18}]$ If BMI is Obesity, Glucose is High, Age is Young, and Skin Thickness is Very Thin, then the decision is 1.

$[R_{19}]$ If BMI is Obesity, Glucose is Low, and Blood Pressure is Hypertension, then the decision is 1.

$[R_{20}]$ If BMI is Obesity, Pregnancy Glucose is Low, and Blood Pressure is Low, then the decision is 1.

$[R_{21}]$ If BMI is Obesity, Glucose is Low, and Blood Pressure is Prehypertension, then the decision is 1.

$[R_{22}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, and Skin Thickness is Normal, then the decision is 1.

$[R_{23}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Thick, Insulin is Low, Pregnancy is Low, and Age is Middle, then the decision is 0.

$[R_{24}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Thick, Insulin is Low, Pregnancy is Low, and Age is Young, then the decision is 1.

$[R_{25}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Thick, Insulin is Low, Pregnancy is Normal, and Age is Middle, then the decision is 1.

$[R_{26}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Thick, Insulin is Low, Pregnancy is Normal, and Age is Young, then the decision is 0.

$[R_{27}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Thick, and Insulin is Normal, then the decision is 1.

$[R_{28}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Very Thin, Pregnancies is Low, and Age is Middle, then the decision is 0.

$[R_{29}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Very Thin, Pregnancy is Low, and Age is Young, then the decision is 1.

$[R_{30}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Very Thin, Pregnancy is Normal, and Age is Middle, then the decision is 1.

$[R_{31}]$ If BMI is Obesity, Glucose is Low, Blood Pressure is Normal, Skin Thickness is Very Thin, Pregnancy is Normal, and Age is Young, then the decision is 0.

$[R_{32}]$ If BMI is Obesity, Glucose is Normal, and Blood Pressure is Hypertension, then the decision is 1.

$[R_{33}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Low, and Pregnancy is High, then the decision is 1.

$[R_{34}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Low, and Pregnancy is Normal, then the decision is 0.

$[R_{35}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Low, Pregnancy is Low, and Age is Middle, then the decision is 0.

$[R_{36}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Low, Pregnancy is Low, and Age is Young, then the decision is 1.

$[R_{37}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancy is High, and Skin Thickness is Normal, then the decision is 1.

$[R_{38}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancy is High, and Skin Thickness is Thick, then the decision is 0.

$[R_{39}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancy is Low, Age is Middle, and Skin Thickness is Thick, then the decision is 0.

$[R_{40}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancy is Low, Age is Middle, and Skin Thickness is Very Thin, then the decision is 1.

$[R_{41}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancies is Low, and Age is Young, then the decision is 1.

$[R_{42}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancies is Normal, and Insulin is Normal, then the decision is 0.

$[R_{43}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancies is Normal, Insulin is Low, and Skin Thickness is Normal, then the decision is 1.

$[R_{44}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancies is Normal, Insulin is Low, and Skin Thickness is Thick, then the decision is 0.

$[R_{45}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Normal, Pregnancies is Normal, Insulin is Low, and Skin Thickness is Very Thin, then the decision is 0.

$[R_{46}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancies is High, Skin Thickness is Thick, and Insulin is Low, then the decision is 0.

$[R_{47}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancies is High, Skin Thickness is Thick, and Insulin is Normal, then the decision is 1.

$[R_{48}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancies is High, and Skin Thickness is Very Thin, then the decision is 1.

$[R_{49}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancies is Low, Skin Thickness is Normal, and Insulin is Low, then the decision is 0.

$[R_{50}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancy is Low, and Skin Thickness is Thick, then the decision is 1.

$[R_{51}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancies is Low, and Skin Thickness is Very Thin, then the decision is 1.

$[R_{52}]$ If BMI is Obesity, Glucose is Normal, Blood Pressure is Prehypertension, and Pregnancies is Normal, then the decision is 0.

$[R_{53}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, and Insulin is High, then the decision is 1.

$[R_{54}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, and Insulin is Normal, then the decision is 1.

$[R_{55}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, and Glucose is High, then the decision is 0.

$[R_{56}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, Glucose is Low, and Pregnancy is Low, then the decision is 1.

$[R_{57}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, Glucose is Low, and Pregnancy is Normal, then the decision is 0.

$[R_{58}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, Glucose is Normal, and Pregnancy is Low, then the decision is 1.

$[R_{59}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, Glucose is Normal, and Pregnancy is Normal, then the decision is 0.

$[R_{60}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Middle, Insulin is Low, Glucose is Normal, Pregnancy is Normal, and Blood Pressure is Prehypertension, then the decision is 1.

$[R_{61}]$ If BMI is Overweight, Skin Thickness is Normal, and Age is Old, then the decision is 1.

$[R_{62}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Young, and Blood Pressure is Low, then the decision is 0.

$[R_{63}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Young, and Blood Pressure is Normal, then the decision is 0.

$[R_{64}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Young, Blood Pressure is Prehypertension, and Pregnancy is Low, then the decision is 0.

$[R_{65}]$ If BMI is Overweight, Skin Thickness is Normal, Age is Young, Blood Pressure is Prehypertension, and Pregnancy is Normal, then the decision is 1.

$[R_{66}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is High, Insulin is Low, and Age is Middle, then the decision is 1.

$[R_{67}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is High, Insulin is Low, and Age is Young, then the decision is 0.

$[R_{68}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is High, and Insulin is Normal, then the decision is 0.

$[R_{69}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Low, and Blood Pressure is Hypertension, then the decision is 0.

$[R_{70}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Low, and Blood Pressure is Low, then the decision is 1.

$[R_{71}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Low, and Blood Pressure is Normal, then the decision is 1.

$[R_{72}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Low, Blood Pressure is Prehypertension, and Pregnancy is Low, then the decision is 0.

$[R_{73}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Low, Blood Pressure is Hypertension, and Pregnancy is Normal, then the decision is 1.

$[R_{74}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, and Blood Pressure is Hypertension, then the decision is 1.

$[R_{75}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, and Blood Pressure is Low, then the decision is 0.

$[R_{76}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Normal, and Pregnancy is High, then the decision is 1.

$[R_{77}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Normal, and Pregnancy is Low, then the decision is 0.

$[R_{78}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Normal, and Pregnancy is Normal, then the decision is 0.

$[R_{79}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Prehypertension, and Pregnancy is High, then the decision is 0.

$[R_{80}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Prehypertension, and Pregnancy is Normal, then the decision is 1.

$[R_{81}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancy is Low, and Age is Middle, then the decision is 0.

$[R_{82}]$ If BMI is Overweight, Skin Thickness is Thick, Glucose is Normal, Blood Pressure is Prehypertension, Pregnancy is Low, and Age is Young, then the decision is 1.

$[R_{83}]$ If BMI is Overweight, Skin Thickness is Thin, and Glucose is High, then the decision is 0.

$[R_{84}]$ If BMI is Overweight, Skin Thickness is Thin, and Glucose is Normal, then the decision is 0.

$[R_{85}]$ If BMI is Overweight, Skin Thickness is Thin, Glucose is Low, and Blood Pressure is Low, then the decision is 0.

$[R_{86}]$ If BMI is Overweight, Skin Thickness is Thin, Glucose is Low, and Blood Pressure is Normal, then the decision is 0.

$[R_{87}]$ If BMI is Overweight, Skin Thickness is Thin, Glucose is Low, and Blood Pressure is Prehypertension, then the decision is 1.

$[R_{88}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Hypertension, and Glucose is High, then the decision is 1.

$[R_{89}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is hypertension, and Glucose is Normal, then the decision is 0.

$[R_{90}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Low, and Age is Middle, then the decision is 0.

$[R_{91}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Low, and Age is Young, then the decision is 1.

$[R_{92}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, and Pregnancy is High, then the decision is 0.

$[R_{93}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is High, and Age is Young, then the decision is 1.

$[R_{94}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is High, and Age is Middle, then the decision is 0.

$[R_{95}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is High, and Age is Young, then the decision is 1.

$[R_{96}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Low, and Age is Middle, then the decision is 1.

$[R_{97}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Low, Age is Young, and Glucose is High, then the decision is 1.

$[R_{98}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Low, Age is Young, and Glucose is Low, then the decision is 0.

$[R_{99}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Low, Age is Young, and Glucose is Normal, then the decision is 0.

$[R_{100}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Normal, Age is Middle, and Glucose is High, then the decision is 1.

$[R_{101}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Normal, Age is Middle, and Glucose is Low, then the decision is 1.

$[R_{102}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Normal, Pregnancy is Normal, Age is Middle, and Glucose is Normal, then the decision is 1.

$[R_{103}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Prehypertension, and Glucose is High, then the decision is 1.

$[R_{104}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Prehypertension, and Glucose is Low, then the decision is 0.

$[R_{105}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Prehypertension, Glucose is Normal, and Age is Middle, then the decision is 0.

$[R_{106}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Prehypertension, Glucose is Normal, and Age is Young, then the decision is 0.

$[R_{107}]$ If BMI is Overweight, Skin Thickness is Very Thin, Blood Pressure is Prehypertension, Glucose is Normal, Age is Young, and Pregnancy is Normal, then the decision is 1.

$[R_{108}]$ If BMI is Underweight, and Blood Pressure is Hypertension, then the decision is 1.

$[R_{109}]$ If BMI is Underweight, and Blood Pressure is Prehypertension, then the decision is 0.

$[R_{110}]$ If BMI is Underweight and Low Blood Pressure, then the decision is 0.

The first and second rules are examples of decisions that make predictions of 0 and 1, respectively; so are the 106th and 107th rules. In the first rule, if someone has a normal BMI, high Glucose levels, high Pregnancy, and Normal Skin Thickness, then it is predicted that their status is not diabetic. In the 106th rule, if someone has an Overweight BMI, Very Thin Skin Thickness, Blood Pressure type, Prehypertension, Normal Glucose levels, and Young Age, then it is predicted that their status is not diabetic. In the second rule, if someone has a normal BMI, high Glucose levels, high Pregnancy, and Thick Skin Thickness, then it is predicted that their status is diabetic. Likewise, in the 107th rule, if someone has an Overweight BMI, Very Thin Skin Thickness, Blood Pressure type, Prehypertension, Normal Glucose levels, Young Age, and Normal Pregnancies, then it is predicted that their status is diabetic.

The confusion matrix for the first iteration of FDTID3-5 with all (8) predictor variables is presented in Table 15.

The confusion matrix shows that True Positive prediction (TP) is 22, True Negative prediction (TN) is 95, false positive (FP) and false negative (FN) predictions are 32 and 5, respectively. Because FP and FN are not zero, or there are wrong predictions, to class 0 and class 1, none of the metrics have a value of 100.

As in CHD modeling, in the other four iterations, the significant variables at the root node to the second node consist of the same variables. However, in the DMD model, there are six variables. They are BMI, Glucose, Skin Thickness, Blood Pressure, Pregnancy, and Age. The last two variables are not found at the root and first nodes. This pattern is also found in the FDTID3 model with five other fuzzy membership function combinations. Therefore, in this research, the FDTID3-1 to FDTID3-6 models for CMD prediction were also built using six and two variables in addition to the complete predictor variables.

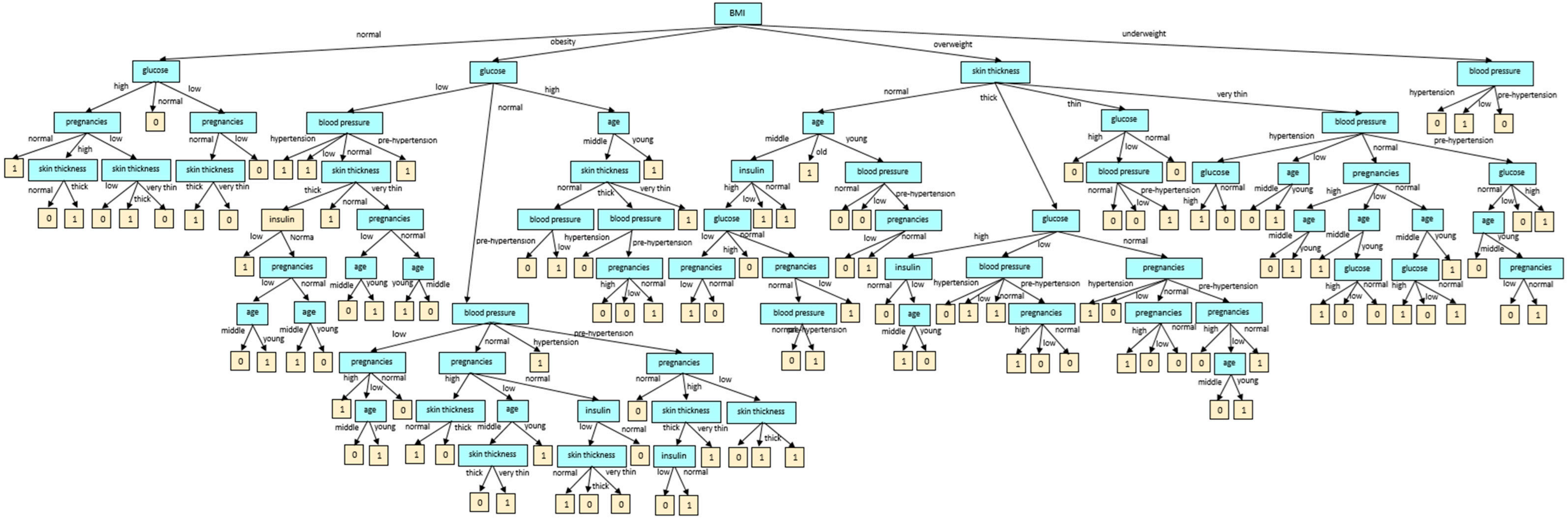
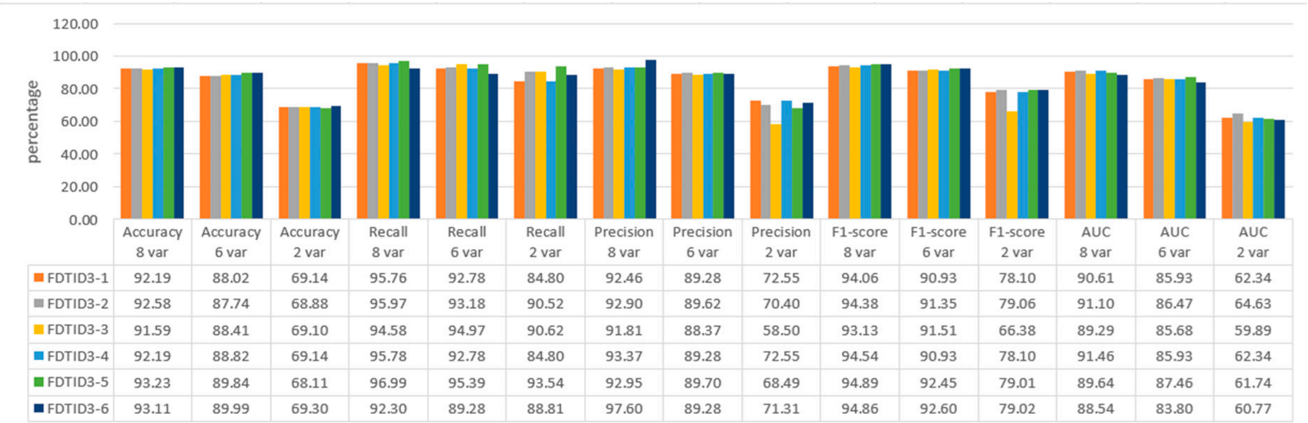


Figure 13. The first iteration of FDT1 final decision tree structure for DMD status prediction.

Table 15. Confusion matrix of DM status prediction of the first iterations of FDTID3-5 with all predictor variables.

		Prediction of DM Status		Sum
The fact of DM Status	Yes	22	32	54
	No	5	95	100
	Sum	27	127	154

The evaluation of the FDTID3-1 to FDTID3-6 prediction models, each based on the number of predictor variables, is presented in Figure 14. These measures are the iteration averages of the 5-fold cross-validation.



	Accuracy 8 var	Accuracy 6 var	Accuracy 2 var	Recall 8 var	Recall 6 var	Recall 2 var	Precision 8 var	Precision 6 var	Precision 2 var	F1-score 8 var	F1-score 6 var	F1-score 2 var	AUC 8 var	AUC 6 var	AUC 2 var
FDTID3-1	92.19	88.02	69.14	95.76	92.78	84.80	92.46	89.28	72.55	94.06	90.93	78.10	90.61	85.93	62.34
FDTID3-2	92.58	87.74	68.88	95.97	93.18	90.52	92.90	89.62	70.40	94.38	91.35	79.06	91.10	86.47	64.63
FDTID3-3	91.59	88.41	69.10	94.58	94.97	90.62	91.81	88.37	58.50	93.13	91.51	66.38	89.29	85.68	59.89
FDTID3-4	92.19	88.82	69.14	95.78	92.78	84.80	93.37	89.28	72.55	94.54	90.93	78.10	91.46	85.93	62.34
FDTID3-5	93.23	89.84	68.11	96.99	95.39	93.54	92.95	89.70	68.49	94.89	92.45	79.01	89.64	87.46	61.74
FDTID3-6	93.11	89.99	69.30	92.30	89.28	88.81	97.60	89.28	71.31	94.86	92.60	79.02	88.54	83.80	60.77

Figure 14. Performance of six FDTID3 prediction models of DMD based on 5-fold cross-validation.

Almost all FDTID3 models have higher metric sizes than DTID3 models that use discretization with crisp sets. This information indicates that almost all combinations of fuzzy membership functions proposed for discretization in the DTID3 method suit the DMD. The model with complete variables (8 variables) is the model that has the highest performance compared to the performance of the other two models, both models with 2 variables and 6 variables. This information also indicates that the FDTID3 model with eight variables best predicts DM status.

Furthermore, a comparison of the performance of all FDTID3 models with DTID3 involving all predictor variables in predicting DMD status is summarized in Table 16. The value of the metrics of each model is the average of the 5-fold performance.

Table 16. The prediction performance of CHD status.

Fuzzy Membership Functions Combination	Prediction Performance Metric (%)				
	Accuracy	Recall	Precision	F1-Score	AUC
DTID3	91.54	92.91	91.19	91.95	87.68
FDTID3-1	92.19	95.76	92.46	94.06	90.61
FDTID3-2	92.58	95.97	92.90	94.38	91.10
FDTID3-3	91.59	94.58	91.81	93.13	89.29
FDTID3-4	92.19	95.78	93.37	94.54	91.46
FDTID3-5	93.23	96.99	92.95	94.89	89.64
FDTID3-6	93.11	92.30	97.60	94.86	88.54

A high recall value is frequently regarded as superior to a high precision value in predicting disease states, such as DM. This is due to the assumption that a model is more effective in predicting a patient's status as positive-sick. Nevertheless, if the patient's status is negative (healthy), the patient's status is predicted to be healthy despite the patient's status being positive. Nevertheless, the accuracy of disease predictions is not inherently improved by more false positives than false negatives. An individual who is predicted to be positive when the patient is negative may cause the patient to experience tension or other excessive responses. A higher F1 score is more advantageous for a prediction metric because it balances False Positives and False Negatives. All the proposed FDTID3 models outperform the DTID3 model. The FDTID3-5 model has the three highest metrics: accuracy, Recall, and F1-score. The number recorded the FDTID3-5 model as the FDTID3 model with the highest metric value compared to other FDTID3 models, mainly because the F1-score it has is the highest. Therefore, the FDTID3-5 method with eight predictor variables (complete) is the best model for predicting DM status. This fact informs that most of the first and third categories in each variable tend to have decreasing sigmoid and increasing sigmoid functions rather than decreasing linear and increasing linear. Next, the second category tends to have a beta function rather than a triangular or pi function if the other two categories are sigmoid functions.

The ANOVA for the seven models is depicted in Table 17 determine whether the performance of the seven proposed models is distinct from that of the other models. The performance of these seven proposed models can be compared by employing Monte Carlo resampling [35]. At 5% significance levels, the ANOVA indicates that the seven proposed models exhibit discrepancies in at least one average performance metric for accuracy, precision, recall, F1-score, and AUC.

Table 17. ANOVA of the proposed model for the DMD.

Metrics	Source of Var.	Sum of Squares	Mean Squares	F	*p*-Value	F-Criteria
Accuracy	between	177.41	29.57	1107.70	3.23×10^{-224}	2.12
	within	9.34	0.03			
Recall	between	488.61	81.43	258.53	2.51×10^{-125}	
	within	110.25	0.31			
Precision	between	212.58	35.43	385.79	6.21×10^{-151}	
	within	32.14	0.09			
F1-score	between	97.40	16.23	391.47	6.74×10^{-152}	
	within	14.51	0.04			
AUC	between	251.12	41.85	418.52	2.50×10^{-156}	
	within	35.00	0.10			

Moreover, which pairs of models perform significantly differently and whether the classification performance metrics of DTID3 have increased when using the proposed FDTID3 models. The post hoc test with a 5% significance level using the Tukey–Kramer is given in Table 18.

Most model pairs have an absolute mean difference (AMD) that exceeds each metric's Q-critical value, namely 0.1, 0.35, 0.19, 0.13, and 0.19. None of the FDTID3 and DTID3 model pairs significantly differ on all five metrics. One model pair has at least three metrics that are significantly different. This fact also informs that the six FDTID3 models also have significantly improved performance from DTID3. Given that the six FDTID3 models have different fuzzy membership functions, it can be concluded that the performance of the proposed FDTID3 model depends on the fuzzy membership function used. Our hypothesis that the seven DTID3 models built differ at least in one metric has been proven, and the performance of the six FDTID3 models is better than the performance of the DTID3 model built using the concept of crisp set membership.

This evidence shows that the novelty of this study is that the performance of the DTID3 model was built using the concept of crisp set membership, which has been successfully improved by discretizing the continuous type of predictor variables using the concept of fuzzy set membership (FDTID3). However, the performance of the FDTID3 model is also influenced by the combination of fuzzy membership functions, including the number of categories in each predictor variable, which is the initial basis, where this study uses expert justification.

Table 18. Tukey–Kramer Test of the proposed model for the DMD.

Comparison Model	Absolute Mean Difference				
	Accuracy	Recall	Precision	F1-Score	AUC
FDTID3-1 vs. FDTID3-2	1.37	4.00	1.03	1.30	1.86
FDTID3-1 vs. FDTID3-3	1.71	1.32	1.12	1.21	1.22
FDTID3-1 vs. FDTID3-4	1.32	1.42	0.61	0.98	2.19
FDTID3-1 vs. FDTID3-5	2.09	1.73	1.32	1.51	0.40
FDTID3-1 vs. FDTID3-6	0.37	0.51	0.17	0.33	0.05
FDTID3-1 vs. DTID3	0.58	1.58	0.42	0.49	1.64
FDTID3-2 vs. FDTID3-3	0.34	2.68	2.14	0.08	1.00
FDTID3-2 vs. FDTID3-4	0.05	2.58	1.64	0.32	1.98
FDTID3-2 vs. FDTID3-5	0.72	2.27	2.35	0.22	0.19
FDTID3-2 vs. FDTID3-6	1.00	3.49	1.20	0.97	0.17
FDTID3-2 vs. DTID3	0.79	2.42	0.61	0.81	0.64
FDTID3-3 vs. FDTID3-4	0.39	0.10	0.50	0.23	0.34
FDTID3-3 vs. FDTID3-5	0.37	0.41	0.20	0.30	1.45
FDTID3-3 vs. FDTID3-6	1.34	0.82	0.94	0.89	1.81
FDTID3-3 vs. DTID3	1.14	0.26	1.53	0.72	0.98
FDTID3-4 vs. FDTID3-5	0.76	0.31	0.71	0.53	0.81
FDTID3-4 vs. FDTID3-6	0.95	0.92	0.44	0.65	1.17
FDTID3-4 vs. DTID3	0.75	0.16	1.03	0.49	1.79
FDTID3-5 vs. FDTID3-6	1.72	1.23	1.15	1.19	2.15
FDTID3-5 vs. DTID3	1.51	0.15	1.74	1.02	1.79
FDTID3-6 vs. DTID3	1.72	1.23	1.15	1.19	0.22

3.3. Model Performance Comparison with Other Research

3.3.1. Coronary Heart Disease

A comparison of model performance using our proposed method on the CHD dataset is presented in Table 19. The comparative research generally proposes some models, but we present only the best model from each research result in the table. The researchers tried to improve the prediction model performance by carrying out various techniques such as rescaling predictor variables [23], variable selection with Relief and Least Absolute Shrinkage Selection Operator (LASSO) [24], variable selection with Logistic Chaos Honey Badger (LCHB) algorithm [26], ensemble technique [20,21,56], balancing class distributions [19], and discretizing the predictor variables using the concept of crisp sets [35].

Considering the importance of recall and F1-score in predicting disease cases, including CHD, based on the performance metrics presented in Table 16, it can be concluded that the best model performance in predicting CHD status is our proposed model that uses the FDTID3-4. It was followed by the DTID3 method with discretization using the crisp set concept [33] and Random Forest (RF) [21], respectively. Furthermore, although the difference in performance metric values between our proposed method and [33] using discretization using a crisp set is tiny, this work has shown, with statistical tests, that the proposed models' performance is significantly different. Further exploring the combination of fuzzy membership functions in discretizing predictor variables could be one way to obtain better model performance. The results of this study also show that involving only selected variables does not always provide better prediction performance.

3.3.2. Diabetes Mellitus Disease

The challenge in any research that predicts the disease status of patients is to obtain satisfactory performance so that the disease can be anticipated early to reduce health costs and improve the sufferer's quality of life. Similarly to the attempts to predict CHD status, numerous attempts have been made to improve the performance of DM prediction models. These include implementing ensemble techniques [22,57], using QDA by violating the assumption that the distribution of predictor variables is Gaussian [58], discarding data with zero value [27], imputing zero-value data with the mean [28], and prediction results using SVM [27], balanced class distribution [28] reducing false classification using DTID3 [27], transforming or discretizing data utilizing crisp sets and the fuzzy sets (our proposed method). Table 17 compares the model performance achieved using the proposed method for the DMD. The comparative studies generally propose several models, but we present in Table 20 only the best model from each research result. Based on the performance metrics in the table, it can be inferred that our proposed model, which employs the FDTID3-5, has the highest accuracy in predicting DM status, given the significance of recall and F1-score in predicting disease cases, including DM. It was followed by the NB method with discretization using the crisp set concept [46] and Gaussian Process (GP) [59], respectively.

Furthermore, although the difference in performance metric values between our proposed method and [46], which uses discretization using a crisp set, is tiny, this work has shown with statistical tests that the performance of the proposed models, including those using discretization using crisp set is significantly different. Further exploring the combination of fuzzy membership functions in discretizing predictor variables could be one way to obtain better model performance, especially on the DMD. These results also show that zero-value data does not always have to be treated, either discarded or imputed with specific values, to improve the prediction model's performance, as long as the prediction method can process the value. Likewise, related to the distribution of classes whose comparison ratio is not too unequal, it does not always have to be balanced.

Table 19. The comparison of model performance (%) of CHD status.

No.	Research	The Best Prediction Method	Validation Method	Accuracy	Recall	Precision	F1-Score	AUC
1	Chowdary et al. [46]	Ensemble of LR, RF, GNB, NNR, KNN	Hold out 67:33	87.00	94.00	91.60	88.00	-
2	Kresnawati et al. [33]	DTID3	10-fold CV	99.63	100.00	99.23	99.61	99.67
3	Hassan et al. [21]	RF	Hold out 70:30	96.28	95.37	96.28	96.28	-
4	Hossen [19]	LR	Hold out 80:20	95.00	95.00	-	-	-
5	Kanwal et al. [24]	SVM with LASSO	Hold out 80:20	85.19	80.77	-	-	-
6	Chandrasekhar and Peddakrishna, 2023 [56]	Ensemble of RF, KNN, LR, NB, GB, AB, SVE	5-fold CV	90.00	89.00	-	-	-
7	Patil and Bhosale, 2023 [23]	FCM-based NN with feature scaling	Hold out 70:30	98.78	-	-	-	-
8	Karthikeyini et al., 2023 [26]	DGRU with LCHB	-	95.15	91.48	92.26	92.21	-
9	Femina and Sudheep, 2020 [36]	Linguistic Fuzzy NB Classifier (LFNBC)	Hold out 90:10	91.30	92.68	-	-	91.44
10	Proposed Method	FDTID3-4	5-fold CV	99.67	100.00	99.29	99.64	99.70

Table 20. The comparison of model performance (%) of DM status.

No.	Research	Zero-Value Data	Balance Class	The Best Prediction Method	Validation Method	Recall	F1-Score	AUC	Accuracy	Precision
1	Maniruzzaman et al., 2017 [59]	impute by median	no special treatment	GP	10-fold CV	91.79	88.22	-	81.97	84.91
2	Shanmugapriya et al., 2017 [37]	no special treatment	no special treatment	SVM	Hold out 75:25	58.90	-	-	73.82	-
3	Tigga and Garg, 2020 [57]	no special treatment	no special treatment	SVM	10-fold CV	77.50	81.30	77.10	74.40	85.60
4	Resti et al., 2021 [46]	no special treatment	no special treatment	NB	5-fold CV	94.48	94.15	-	95.83	93.82
5	Tasin et al., 2022 [22]	impute by mean (for skin thickness and BMI) and impute by XGB (for others)	balanced using ADASYN	XGBoost	Hold out 80:20	80.00	81.00	-	88.50	82.00
6	Kresnawati et al., 2023 [58]	no special treatment	no special treatment	QDA	Hold out 70:30	69.23	81.82	84.62	98.27	100.00
7	Binerbia, 2022 [28]	impute by mean	no special treatment	SVM	Hold out 80:20	86.00	-	-	80.00	75.00
8	Palanivinayagam and Damasevicius, 2023 [27]	impute by SVM	no special treatment	SVM	10-fold CV	88.23	85.71	-	94.89	83.33
9	Proposed Method	no special treatment	no special treatment	FDTID3-5	5-fold CV	96.99	94.89	89.64	93.23	92.95

4. Conclusions

This study has predicted the status of degenerative diseases, coronary heart disease, and diabetes mellitus by building seven models of DTID3, respectively. One DTID3 model uses the concept of crisp set membership, and six DTID3 models use the concept of fuzzy set membership with final membership selection rules used as maximum value (FDTID3). The hypothesis that the performance of the seven proposed models differs at least in one metric and that the performance of the FDTID3 models is higher than the DTID3 model discretized using the concept of crisp sets has been proven. The evidence shows that the novelty of this study is that the performance of the DTID3 model built using the concept of crisp set membership has been successfully improved by discretizing the continuous type of predictor variables using the concept of fuzzy set membership (FDTID3). However, the performance of the FDTID3 model is also influenced by the combination of fuzzy membership functions, including the number of categories in each predictor variable, which is the initial basis, where this study uses expert justification. We also note that involving significant variable selection, treating zero-value data, and balancing class distributions does not always perform better than original model which discretizing continuous predictor variables.

Author Contributions: Conceptualization, E.S.K. and Y.R.; methodology, E.S.K. and Y.R.; software, E.S.K. and Y.R.; validation, E.S.K. and B.S.; formal analysis, E.S.K., B.S. and Y.R.; investigation, E.S.K. and Y.R.; resources, E.S.K. and Y.R.; data curation, E.S.K. and Y.R.; writing—original draft preparation, E.S.K. and Y.R.; writing—review and editing, B.S.; visualization, B.S. and Y.R.; supervision, Y.R.; project administration, E.S.K.; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from science, technology, and arts research, FMIPA, Universitas Sriwijaya, with no.0268/UN9. FMIPA/TU.SK/2023.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors thank Universitas Sriwijaya for supporting this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kuo, N.I.; Jorm, L.; Barbieri, S. Synthetic health-related longitudinal data with mixed-type variables generated using diffusion models. *arXiv* **2023**. [CrossRef]
2. Nezhad, S.N.; Zahedi, M.H.; Farahani, E. Detecting diseases in medical prescriptions using data mining methods. *BioData Min.* **2022**, *15*, 29. [CrossRef] [PubMed]
3. Kee, O.T.; Harun, H.; Mustafa, N.; Murad, N.A.A.; Chin, S.F.; Jaafar, R.; Abdullah, N. Cardiovascular complications in a diabetes prediction model using machine learning: A systematic review. *Cardiovasc. Diabetol.* **2023**, *22*, 13. [CrossRef] [PubMed]
4. Abdalrada, A.S.; Abawajy, J.; Al-Quraishi, T.; Islam, S.M.S. Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study. *J. Diabetes Metab. Disord.* **2022**, *21*, 251–261. [CrossRef]
5. Eadie, M.J. The Australian Journal of Physiotherapy degenerative disease affecting the nervous system. *Aust. J. Physiother.* **1974**, *20*, 20–22. [CrossRef]
6. Batista, P.; Pereira, A. Quality of life in patients with neurodegenerative diseases. Imedpub journals quality of life in patients with neurodegenerative diseases. *J. Neurol. Neurosci.* **2016**, *7*, 74. [CrossRef]
7. Harahap, J.; Andayani, L.S. Screening of Degenerative Diseases and Quality of Life among Elderly People in Posyandu Lansia Medan Amplas. In Proceedings of the 5th Annual International Conference Syiah Kuala University, Banda Aceh, Indonesia, 9 September 2015.
8. Barendregt, J.J.M. Degenerative Disease in an Aging Population Models and Conjectures. Ph.D. Thesis, The Department of Public Health of Erasmus Universiteit, Rotterdam, The Netherlands, 1998.
9. Di Renzo, L.; Gualtieri, P.; Frank, G.; De Lorenzo, A. Nutrition for prevention and control of chronic degenerative diseases and COVID-19. *Nutrients* **2023**, *15*, 2253. [CrossRef]
10. Livingston, K.A.; Freeman, K.J.; Friedman, S.M.; Stout, R.W.; Lianov, L.S.; Drozek, D.; Shallow, J.; Shurney, D.; Patel, P.M.; Campbell, T.M.; et al. Lifestyle medicine and economics: A proposal for research priorities informed by a case series of disease reversal. *J. Environ. Res. Public Health* **2021**, *18*, 11364. [CrossRef]
11. Nelwan, E.J.; Widjajanto, E.; Andarini, S.; Djati, M.S. Modified risk factors for coronary heart disease (CHD) in Minahasa ethnic group from Manado city Indonesia. *J. Exp. Life Sci.* **2016**, *6*, 88–94. [CrossRef]

12. Di Cesare, M.; Bixby, H.; Gaziano, T.; Hadeed, L.; Kabudula, C.; McGhie, D.V.; Mwangi, J.; Pervan, B.; Perel, P.; Piñeiro, D.; et al. *World Heart Report 2023 Confronting the World's Number One Killer*; World Heart Federation: Geneva, Switzerland, 2023.
13. Antini, C.; Caixeta, R.; Luciani, S.; Hennis, A.J.M. Diabetes mortality: Trends and multi-country analysis of the Americas from 2000 to 2019. *Int. J. Epidemiol.* **2024**, *53*, dyad182. [CrossRef]
14. WHO. *Global Report on Diabetes*; WHO Library Cataloguing in Publication Data: Lyon, France, 2016.
15. IDF. *Diabetes Voice*; IDF: Brussels, Belgium, 2017; Volume 64.
16. Abdollahi, J.; Moghaddam, B.N.; Parvar, E. Improving diabetes diagnosis in smart health using genetic-based ensemble learning algorithm approach to IoT infrastructure. *Future Gener. Distrib. Syst. J.* **2019**, *1*, 26–33.
17. Cavan, D.; Makaroff, L.; Fernandes, J.D.R. *Cost-Effective Solutions for the Prevention of Type 2 Diabetes*; IDF: Brussels, Belgium, 2016.
18. WHO. *World Health Statistics Overview 2019*; WHO: Geneva, Switzerland, 2019.
19. Hossen, M.K. Heart disease prediction using machine learning techniques. *Am. J. Comput. Sci. Technol.* **2022**, *5*, 146–154. [CrossRef]
20. Chowdary, G.J.; Suganya, G.; Mariappan, P. Predicting the presence of coronary heart disease using machine learning classifiers. *J. Crit. Rev.* **2020**, *7*, 1865–1875.
21. Hassan, C.A.U.; Iqbal, J.; Irfan, R.; Hussain, S.; Algami, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors* **2022**, *22*, 7227. [CrossRef] [PubMed]
22. Tasin, I.; Nabil, T.U.; Islam, S.; Khan, R. Diabetes prediction using machine learning and explainable AI. *Healthc. Technol. Lett.* **2023**, *10*, 1–10. [CrossRef]
23. Patil, S.; Bhosale, S. Improving cardiovascular disease prognosis using outlier detection and hyperparameter optimization of machine learning models. *Rev. d'Intell. Artif.* **2023**, *37*, 1069–1080. [CrossRef]
24. Kanwal, A.; Ahmad, K.T.; Abid, M.K.; Aslam, N. Detection of heart disease using supervised machine learning. *Vfast Trans. Softw. Eng.* **2022**, *6246*, 58–70. [CrossRef]
25. Selvan, S.; Varadhaganapathy, S. Deep learning based cardiovascular disease risk factor prediction among type 2 diabetes mellitus patients. *Inf. Technol. Control* **2023**, *52*, 215–227. [CrossRef]
26. Karthikeyini, S.; Vidhya, G.; Vetriselvi, T.; Deepa, K. Heart disease prognosis using D-GRU with logistic chaos honey badger optimization in IOMT framework. *Inf. Technol. Control* **2023**, *52*, 367–380. [CrossRef]
27. Palanivinayagam, A.; Damaševičius, R. Effective handling of missing values in datasets for classification using machine learning methods. *Information* **2023**, *14*, 92. [CrossRef]
28. Benarbia, M. A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women. Master's Thesis, City University of New York, NY, USA, 2022.
29. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe, CA, USA, 9–12 July 1995. [CrossRef]
30. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: Cham, Switzerland, 2015; Volume 72.
31. Roy, A.; Pal, S.K. Fuzzy discretization of feature space for a rough set classifier. *Pattern Recognit. Lett.* **2003**, *24*, 895–902. [CrossRef]
32. Resti, Y. Credit Risk-Type Classification using Statistical Learning. In Proceedings of the 3rd Conference on Fundamental and Applied Science for Advanced Technology Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 22 January 2022. [CrossRef]
33. Kresnawati, E.S.; Resti, Y.; Suprihatin, B.; Kurniawan, M.R.; Amanda, W.A. Coronary artery disease prediction using decision trees and multinomial naïve bayes with k-fold cross validation. *Inomatika* **2021**, *3*, 174–189. [CrossRef]
34. Resti, Y.; Irsan, C.; Amini, M.; Yani, I.; Passarella, R. Performance improvement of decision tree model using fuzzy membership function for classification of corn plant diseases and pests. *Sci. Technol. Indones.* **2022**, *7*, 284–290. [CrossRef]
35. Resti, Y.; Irsan, C.; Neardiaty, A.; Annabila, C.; Yani, I. Fuzzy discretization on the multinomial naïve Bayes method for modeling multiclass classification of corn plant diseases and pests. *Mathematics* **2023**, *11*, 1761. [CrossRef]
36. Femina, B.T.; Sudheep, E.M. A novel fuzzy linguistic fusion approach to naive Bayes classifier for decision-making applications. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2020**, *10*, 1889–1897. [CrossRef]
37. Shanmugapriya, M.; Nehemiah, H.K.; Bhuvaneswaran, R.S.; Arputharaj, K.; Sweetlin, J.D. Fuzzy discretization based classification of medical data. *Res. J. Appl. Sci. Eng. Technol.* **2017**, *14*, 291–298. [CrossRef]
38. Tutuncu, G.Y.; Kayaalp, N. An aggregated fuzzy naive Bayes data classifier. *J. Comput. Appl. Math.* **2019**, *286*, 17–27. [CrossRef]
39. Algehyne, E.A.; Jibril, M.L.; Algehainy, N.A.; Alamri, O.A.; Alzahrani, A.K. Fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia. *Big Data Cogn. Comput.* **2022**, *6*, 13. [CrossRef]
40. Altay, A.; Cinar, D. Fuzzy decision trees. In *Fuzzy Statistical Decision-Making*; Springer International Publisher: Cham, Switzerland, 2016. [CrossRef]
41. Araniba, L.A.Q. Learning Fuzzy Logic from Examples. Master's Thesis, Ohio University, Athens, OH, USA, 1994.
42. Resti, Y.; Burlian, F.; Yani, I.; Zayanti, D.A.; Sari, I.M. Improved the cans waste classification rate of naive Bayes using fuzzy approach. *Sci. Technol. Indones.* **2020**, *5*, 75–78. [CrossRef]
43. Fernandez, S.; Ito, T.; Cruz-Piris, L.; Marsa-Maestre, I. Fuzzy ontology-based system for driver behavior classification. *Sensor* **2022**, *22*, 7954. [CrossRef]
44. Kaggle. Available online: https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset/data (accessed on 17 January 2024).

45. Kaggle. Available online: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed on 17 January 2024).
46. Resti, Y.; Kresnawati, E.S.; Dewi, N.R.; Zayanti, D.A.; Eliyati, N. Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive Bayes, discriminant analysis, and logistic regression. *Sci. Technol. Indones.* **2021**, *6*, 96–104. [CrossRef]
47. Lee, C.F.; Tzeng, G.H.; Wang, S.Y. A new application of fuzzy set theory to the black-scholes option pricing model. *Expert Syst. Appl.* **2005**, *29*, 330–342. [CrossRef]
48. Resti, Y.; Irsan, C.; Putri, M.T.; Yani, I.; Ansyori, A.; Suprihatin, B. Identification of corn plant diseases and pests based on digital images using multinomial naïve Bayes and k-nearest neighbor. *Sci. Technol. Indones.* **2022**, *7*, 29–35. [CrossRef]
49. Bhattacharyya, R.; Mukherjee, S. Fuzzy membership function evaluation by non-linear regression: An algorithmic approach. *Fuzzy Inf. Eng.* **2021**, *12*, 412–434. [CrossRef]
50. Alzoman, R.M.; Alenazi, M.J.F. A comparative study of traffic classification techniques for smart city networks. *Sensors* **2021**, *21*, 4677. [CrossRef]
51. Rutkowski, L. *Flexible Neuro-Fuzzy Systems*; Kluwer Academic Publisher: Boston, FL, USA, 2004.
52. Medasani, S.; Kim, J.; Krishnapuram, R. An overview of membership function generation techniques for pattern recognition. *Int. J. Approx. Reason.* **1998**, *19*, 391–417. [CrossRef]
53. Lantz, B. *Machine Learning with R*; Packt Publishing: Birmingham, UK, 2013; pp. 315–348.
54. Rodríguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 569–575. [CrossRef]
55. Ramasubramanian, K.; Singh, A. *Machine Learning Using R*, 2nd ed.; Apress: Berkeley, CA, USA, 2019. [CrossRef]
56. Chandrasekhar, N.; Peddakrishna, S. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes* **2023**, *11*, 1210. [CrossRef]
57. Tigga, N.P.; Garg, S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Comput. Sci.* **2020**, *167*, 706–716. [CrossRef]
58. Kresnawati, E.S.; Suprihatin, B.; Resti, Y. Diabetes Mellitus Diagnosis Using The Prediction Model of Discriminant Analysis. In Proceedings of the AIP Conference Proceedings of Annual Conference on Science and Technology Research, Palembang, Indonesia, 24 August 2021. [CrossRef]
59. Maniruzzaman, M.; Kumar, N.; Abedin, M.M.; Islam, M.S.; Suri, H.S.; El-Baz, A.S.; Suri, J.S. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Comput. Methods Programs Biomed.* **2017**, *152*, 23–34. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.