CHAPTER 3
RESEARCH METHODOLOGY

3.1 Overview

In this chapter, we focus on the research methodology that is used. The first methodology is about how the data is being prepared and preprocessed. And then the next methodology is about how the two algorithms perform on diabetes prediction and comparing both of them. Figure 3.1 will visualize the used methodology.

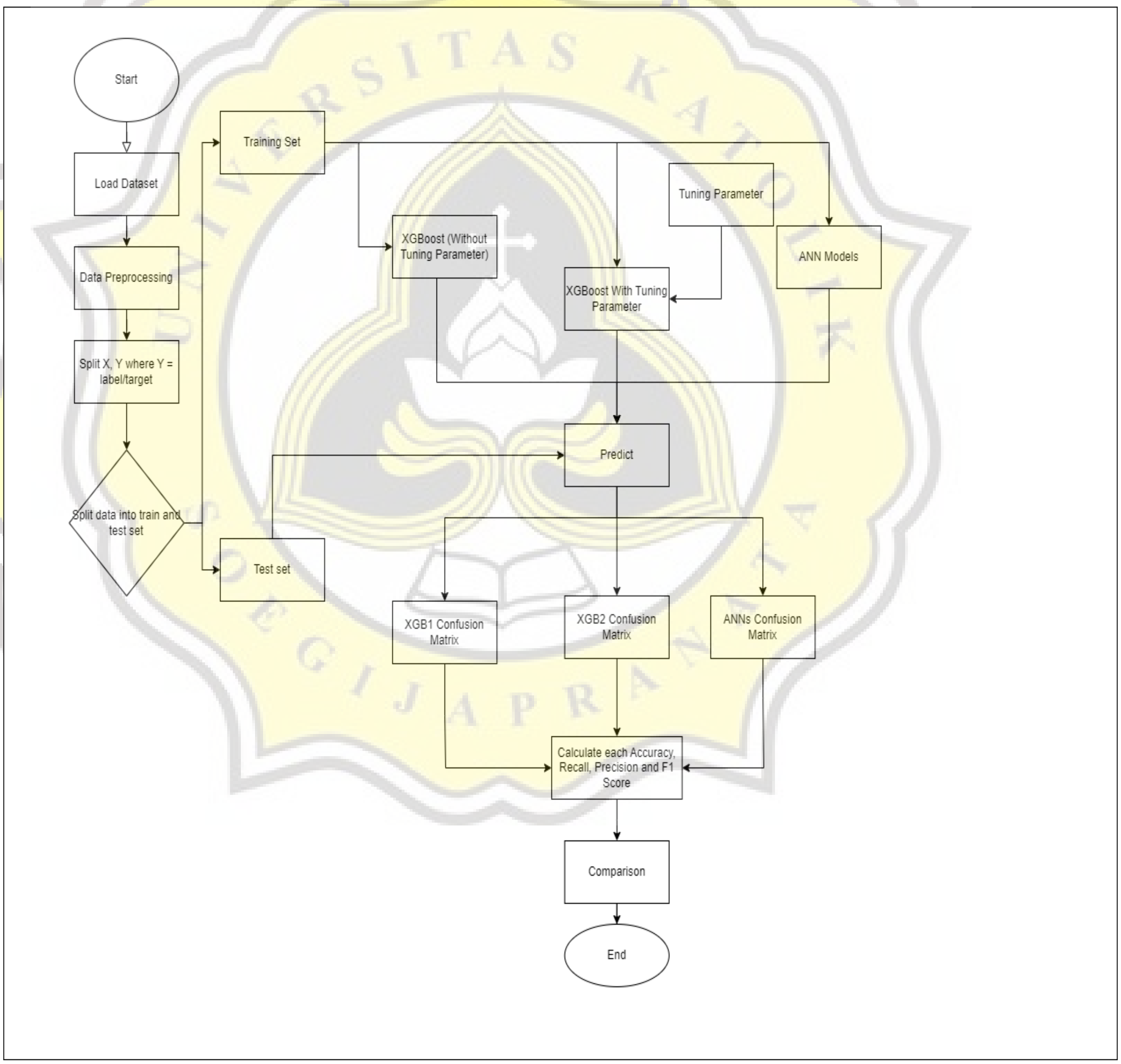
Start
Load Dataset
Data Preprocessing
Split X, Y where Y = label/target
Split data into train and test set
Training Set
Test set
XGBoost (Without Tuning Parameter)
Tuning Parameter
XGBoost With Tuning Parameter
ANN Models
Predict
XGB1 Confusion Matrix
XGB2 Confusion Matrix
ANNs Confusion Matrix
Calculate each Accuracy, Recall, Precision and F1 Score
Comparison
End

Figure 3.1 Flowchart of This Study

3.2 Dataset

The used dataset of this study obtained from kaggle called Diabetes UCI Dataset which has been collected through direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved by a doctor. It has total 520 data and 17 attributes including the ‘class’ attributes where Positive indicates the person had diabetes and Negative doesn’t have diabetes . Table 3.1 shows the attributes information.

Table 3.1 Table Dataset UCI Diabetes Attributes Information

Attributes name	Details
Age	20-65
Sex	Male/Female
Polyuria	Yes/No
Polydipsia	Yes/No
sudden weight loss	Yes/No
weakness	Yes/No
Polyphagia	Yes/No
Genital thrush	Yes/No
visual blurring	Yes/No
Itching	Yes/No
Irritability	Yes/No
delayed healing	Yes/No
partial paresis	Yes/No
muscle stiffness	Yes/No
Alopecia	Yes/No
Obesity	Yes/No
Class	Positive/Negative

3.2.1 Data Preprocessing

Before training the dataset with the algorithms, data preprocessing was performed for the efficient performance of training [2],[4],[6]. It has a total of 520 and 17 attributes. During this preprocessing process, the dataset was checked if it has any missing values and the result is it has no missing values. And one hot encoding was performed where changing male into 1 and female 0, yes into 1 and no 2, and for the class were positive into 1 and negative into 0 [16]. Then splitting the dataset into X and Y where X is the attributes and Y is label or target, where Y is 'Class' that indicates either the person's positive or negative diabetes. And the dataset was divided into two different sets, namely for training and test purposes.

3.3 Artificial Neural Network (ANN)

Artificial Neural Network or ANN is a biological neural network model based machine learning algorithm. It is affected by the information that flows through the network since it learns through the input and output [4]. ANN general structure consists of input layer, hidden layer, and output layer. A process of adjusting the weights of all features by comparing the output value with the actual value in order to improve the precision of the model is the fundamental of ANN. The following is the basic learning process of ANN:

1. First, the perceptron receives the input values and applies the activation function to the output value while initializing the weights close to 0 [8].
2. Second, comparing the output value with the actual value and measuring the error in the prediction using the cost function [8].
3. Lastly, back propagation is applied to the neural network to adjust and update the information of weights to minimize the error in the second step, gradient descent is applied in this process to minimize the cost function. And then repeat from the first step until the const function is minimized [8].

3.4 Extreme Gradient Boosting (XG Boost)

Extreme gradient boosting or XGBoost is a tree boosting based scalable end-to-end algorithm which is used widely by data scientists to achieve state of the art results on many machine learning challenges. It was developed by Tianqi Chen and Carlos Guestrin[11]. It was introduced in 2014 and often used because of its speed , scalability, and effectiveness to solve classification or regression problems [12]. If the result shows high training accuracy, but low test accuracy, it is likely the model was overfitting , so to control the overfitting, a parameter tuning process is needed. Phan et al. study results showed that to demonstrate the superiority of the proposed XG Boost, parameter tuning is applied and it shows that the XGBoost model needs to be modified by the tuning parameter process [13]. The following are the two general way to control overfitting:

First way is by directly controlling the model complexity by tuning “max_depth” or maximum depth of a tree, “min_child_weight” is the minimum sum of weight in each child and “gamma” is the minimum loss reduction that required to make a further partition on a leaf of the tree.

Second way is to make the training robust to noise by adding some randomness, by tuning “subsample” is the ratio of the training instances and “colsample_bytree” is the subsample ratio of columns when constructing each tree. We can also reduce the “eta” or the learning rate but also increase the “num_round” or number of rounds for boosting when you do so.

In this study Bayesian Optimization is used, Hossain et al. shows that Bayesian Optimization is more efficient than Grid Search and Random Search because it can find the optimal combinations of hyperparameters by analyzing the previously tested values, and running the model based on previous tested values is much cheaper than running the whole objective function [14].

3.5 Evaluation Model

The evaluation will be seen by the accuracy rate, precision, recall, and F1 score whether XGBoost can be used or will have better results compared to ANN or not on diabetes prediction.

In this study, the analysis is done by getting the accuracy, precision, recall and F1 score of both proposed algorithms and comparing them to find out which has the better result.

Accuracy is the overall success rate of the model, Figure 3.2 shows the formula of accuracy [2],[6],[15].

$$Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative}$$

Figure 3.2 Accuracy

Precision is defined as the number of true positives compared to positive predictions, precision is shown at Figure 3.3 [6][15].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Figure 3.3 Precision

Recall is to measure the classifier completeness or the sensitivity, Figure 3.4 shows the formula of recall [6],[15].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ \ Negative}$$

Figure 3.4 Recall

F1 score is the weighted average of the precision and recall, Figure 3.5 shows the formula of F1 score [6],[15].

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

Figure 3.5 F1 Score

After getting the accuracy, precision, recall, and f1 score of both algorithms, we compare it with each other to know if XGBoost can be used or has a better result than the ANN on diabetes prediction.