CHAPTER 3
RESEARCH METHODOLOGY

3.1. Journal References

At this stage, a review of various journals related to research and other references is carried out. Such as journals that discussing breast cancer, the Naïve Bayes theorem, and the Wisconsin Database. This research was conducted to complement the concepts and theories that support this research and to collect the data needed for this study.

3.2. Data Collection

The data Wisconsin Breast Cancer Database are extracted from the UCI Machine Learning repository: (http://archive.ics. uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/) [14]. This dataset is collected by Dr. William H. Wolberg at the University of Wisconsin. The amount of data is 699 records with 11 attributes including class attributes. These attributes are displayed in Table 1.

Table 3.1. Dataset Overview

NO	Attributes	Domain	Description
1	Sample Code Number	1 - 10	The attributes as sample ID number
2	Clump Thickness	1 – 10	This attribute determines whether the cell is layered or not. Malignant cells tend to be grouped in multilayer.
3	Uniformity Cell Size	1 – 10	This attribute determines the cell size similarity, because cancer cells have various sizes.
4	Uniformity Cell Shape	1 – 10	This attribute determines the similarity of the cell shape, because cancer cells have various shape and tend to be unnatural.
5	Marginal Adhesion	1 – 10	This attribute determines whether the cell is normal or not. Normal cells tend to stick together, whereas cancer cells don't
6	Single Epithelial Cell Size	1 – 10	This attribute determines whether the epithelial cell tend to enlarge or not.
7	Bare Nuclei	1 – 10	This attribute determines whether the cell is surrounded by cytoplasm or not
8	Bland Chromatin	1 - 10	This attribute determines the texture level of the cell chromatin
9	Normal Nucleoli	1 - 10	This attribute determines the shape of the nucleoli. The higher the value, the more abnormal the nucleoli are.
10	Mitoses	1 - 10	This attribute determines how much the cancer cell divides or reproduces itself. Later will indicate classified as benign or malignant.

11	Class	2 and 4	The are two classes, grade 2 for Benign and grade 4 for Malignant.

3.3. Code

The classification system will be implemented in Python and dataset stores as a .csv file.

3.4. Implementation and Analysis

The medical dataset that will be implemented comes from UCI Machine Learning (*http://archive.ics. uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/*) and uses Jupyter notebook tools. The classification model was built using the Naïve Bayes classification process.

First, the data are extracted from UCI repositories [14] as input data and stored as a .csv file. This WBC dataset contains 699 instances and 11 attributes of which 70% of the data has been retrieved for training purpose and 30% for testing purpose. These testing data are applied on classification method, which detect whether the cell is malignant or benign.

Then next step is *pre-processing* on data. In this dataset, there are 16 missing values and must be eliminated to execute the classification algorithm. In addition, the ID number feature cannot provide useful information for diagnosis, so this feature is omitted from the dataset.

Table 3.2. Dataset After Cleaning

	Breast Cancer Dataset
Preliminary Data	699
Missing Value	16
Final Data	683

In Table 2, it can be seen that the number of data after cleaning is 683. These data are divided into 444 benign data and 239 malignant data.

The following flowchart represents the detection of breast cancer cells using Naïve Bayes as a diagnostic technique.

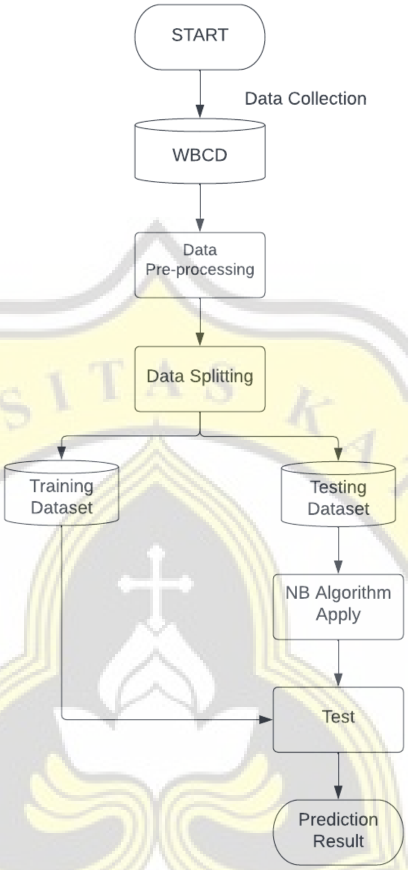

Figure 3.1 Flowchart of The Implementation of The Naive Bayes Method for Breast Cancer Prediction

Then, the resulting dataset is divided into two training and testing datasets. In this step of the proposed method, 70% of the data are randomly selected for training and the remaining 30% are selected for testing.

In the next step, Naïve Bayes are executed on the training dataset to create predictor models. At this stage, after reading the training data, statistical calculations are performed to obtain the average, standard deviation, and probability of each attribute.

After obtaining the mean, and probability then the criteria can be generated. The results of this prediction will be tested on testing dataset to see the accuracy.

3.5. Conclusion and Report Writing

The final result that is expected to be able to classify the type of cancer and will be displayed in the form of a table. The output of the prediction algorithm must be the same as the output of the classification algorithm. This will also verify the results of prediction algorithm