I. Introduction

Stunting is a chronic nutritional problem caused by prolonged inadequate nutrient intake, leading to growth disorders where a child's height is significantly below the age standard[12]. Although the stunting rate in Indonesia has decreased significantly to 21.6% based on www.kemkes.go.id, it remains a major health concern for children, as the World Health Organization (WHO) recommends a prevalence below 20%[4].

One of the best solutions to stunting is prevention to avoid the problem itself. To reduce stunting rates in Indonesia, preventive measures and improved nutrition must be taken before stunting affects a child. One way to avoid it is to monitor the children's growth regularly[13]. Thus, a system to predict the potential for stunting in children is needed[9]. Machine learning offers methods to predict stunting in toddlers, enabling health workers to provide early nutritional guidance. Based on the research in[2], the Ensemble Learning Boosted K-Nearest Neighbor (BK) performs well in predicting stunting conditions. In 2023, the stunting problem was discussed in[2], which used BK to predict the same stunting dataset. The BK method got 98% accuracy as the highest result after balancing the dataset and several iterations of learning. On the other hand, the research in[7] used the Naïve Bayes method to predict the same stunting dataset and got 64.36% accuracy as the highest result.

In order to cope with imbalanced data classification, research in[6] proposes an ensemble algorithm named BPSO-Adaboost-KNN. The main idea of this algorithm is to integrate feature selection and boosting into an ensemble. On the other hand, research in[10] proposes Naïve Bayes classifier as the solution. But Naïve Bayes' effectiveness still needs to be upgraded, so they presented solutions on using Boosting to improve the Gaussian Naïve Bayes algorithm by combining the Naïve Bayes classifier and Adaboost methods[11].

This study seeks to determine the best machine-learning model dedicated to detecting stunting conditions in toddlers between Ensemble Learning Algorithm. Leveraging a dataset sourced from the Bojongsoang Community Health Center, consisting of over 7,500 records of toddlers' physical measurements, the research focuses on training a model capable of providing early warnings. The comparison between two Ensemble Machine Learning models—Boosted K-Nearest Neighbor (BK) and Boosted Naïve Bayes (BN)—is meticulously outlined to identify the most effective model. This research aims to contribute to stunting prevention in Indonesia by deploying machine learning models to provide timely alerts, potentially mitigating the impact of this critical health issue.

II. Methods

A. Research Design

This research starts by reviewing related literature to this research to get the background of the problem and acquire the idea of methods used to solve the problem in this research. The research continues by collecting toddler data from the Bojongsoang Community Health Center. The collected data then needs to be understood first. After understanding each feature, some features are selected to be processed later. The features that have been selected are visualized so they can be easier to explain. Before the method chosen is implemented in the data, the data needs to be cleaned and converted to compatible data types to be processed. The data then needs to be checked to see whether it is balanced. The data will be balanced first if it turns out that the data is an imbalanced dataset[2]. After preprocessing, BK and BN will try to classify the data. The performance of each method will be compared to get the conclusion. To get a better explanation, Fig. 1 will show the flowchart used for the research.

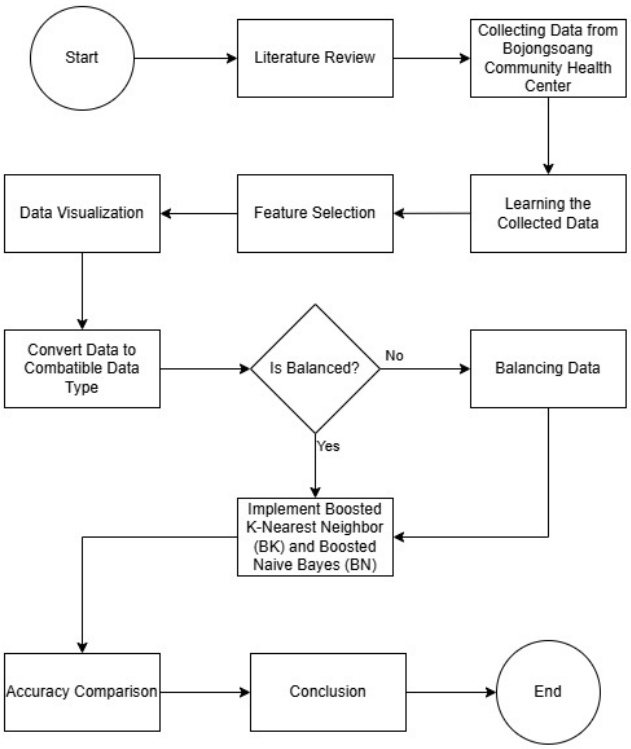

Fig. 1. Research Flowchart

B. K-Nearest Neighbor

K-Nearest Neighbor (KNN) algorithm is one of the most common algorithms used to do classification or regression on data[3]. This algorithm is employed for its simplicity and proven efficacy in classification tasks[14]. The idea of the KNN algorithm is to find similarity in each data, selecting k objects set that are the most similar, and labeling the new data based on the selected k objects set[15]. The similarity between data is determined by calculating the distance between data using Euclidean distance[15]. Euclidean distance from $l = (l_1, \cdots, l_n)$ to $m = (m_1, \cdots, m_n)$ is given as,

$$euc(l, m) = \sqrt{\sum_{i=1}^{n} (l_i - m_i)^2} \tag{1}$$

where n is the number of columns or features in the data and the smaller the distance, the more similar the data are[15]. The KNN algorithm classifies data based on the distance between each unlabeled data and all labeled data in the dataset[2]. The classification is based on KNN (smallest distances), where k is the number of neighbors involved in the voting process[2]. The class label for the test data is determined based on the majority votes[1].