I. Introduction

Eruptive activity that started in June 2018 caused a landslide on Anak Krakatau volcano in December 2018. Volcanic material was released into the sea by the landslide and caused a tsunami with a height of up to 13 meters near Sumatra and Java [1]. At least 437 people died in the natural disaster, making it one of the deadliest volcano-generated tsunamis [1]. The tsunami early warning system deployed in Indonesia was ineffective since it could only detect tsunamis based on earthquake data rather than sea level data [2]. To strengthen the tsunami early warning system, Badan Pengkajian dan Penerapan Teknologi (BPPT) installed the Indonesia Cable Base Tsunameter (InaCBT) [3]. Labuan Bajo and Rokatenda in East Nusa Tenggara (NTT) are two areas where InaCBT has been deployed [3]. The cost of installing InaCBT and developing infrastructure in Labuan Bajo and Rokatenda totaled Rp66 billion [3]. Meanwhile, electronic systems, cables, and testing equipment cost approximately Rp50 billion [3]. These are not small expenses.

In computer science, a study called machine learning aims to make computers learn automatically without explicit programming [4]. Machine learning can be utilized for various purposes, including disaster prevention, for example, developing an effective and inexpensive tsunami early warning system. The utilization of machine learning to detect tsunamis has already been done in previous research. Dilectin and Mercy used KNN to implement tsunami classification for real-time data and the automated detection of novel classes for outliers [5]. The computer vision approach was employed to increase the effectiveness of the Tsunami Warning System by extracting tsunami-like wave footage [6]. Dewi and Diah employed a Bayesian classifier to predict the tsunami potential based on the earthquake data [7]. Francesco, Davide, and Luca developed a new tsunami detection algorithm based on real-time tide removal and band-pass filtering of sea-bed pressure measurements [8].

In previous research, tsunami early warning systems were developed using wave, tidal, and earthquake data. However, other options are required to enhance the system, such as using sea level data as an alternative information source. To

utilize sea level data, the device must be able to detect anomaly signals, especially in distinguishing signals caused by tsunamis and those not caused by tsunamis. Therefore, a more integrated tsunami early warning system based on sea level data that can detect anomalies for tsunami and non-tsunami events is needed.

In this paper, we studied a tsunami early warning system modeling using Extreme Gradient Boosting (XGBoost) machine learning method to classify tsunami and non-tsunami signals based on sea level data. Feature engineering was conducted to obtain influential features in improving the model's performance in classifying tsunami signals. As a study case, we used sea level data obtained from the Inexpensive Device for Sea Level Measurement (IDSL) in Marina Jambu, Banten, Indonesia (latitude: -6.189322, longitude: 105.841088) [2]. The data have been artificially added with tsunami signals from other periods to increase the number of tsunami signals.

The following is the paper's structure: Section II addresses the literature review on the tsunami early warning system, the architecture of XGBoost, and model evaluation metrics. Section III follows with a summary of our study's methodology. In Section IV, we describe the classification results utilizing the XGBoost in four scenarios: lag features, imbalance data handling technique, the gradient of the sea level anomaly as a feature, and times as a feature. Some findings from the paper were concluded in the final section.