I. INTRODUCTION

Ocean waves are an important natural phenomenon that affects offshore and coastal operations. Wave height analysis is essential to ensure navigation safety, optimise marine operations, and reduce the hazards of natural disasters. Precise wave height forecasts can increase production, save fuel and prevent dangerous scenarios. Individuals in marine and coastal cities can benefit from systems capable of forecasting wave patterns around the clock. However, ocean waves' complex and erratic nature makes it challenging to calculate wave height [1].

Several methods exist for precisely forecasting wave heights. Nonetheless, several issues remain unresolved. Numerical approaches cannot deliver intricate wave representations and need substantial processing resources [2]. Conversely, statistical methodologies employing trend analysis of historical data have been utilised for projecting wave heights. However, these statistical approaches are limited to handling complex scenarios and cannot describe complicated nonlinear interactions [3]. Therefore, machine learning becomes a very effective way to numerically and statistically improve wave forecasting models. This approach was chosen because it can optimise large datasets for accurate predictions. Its effectiveness has been proven and is now used in modern forecasting methods [4]. For example, Anggraeni et al. [2] Conducted research comparing the AdaBoost and XGBoost methodologies for forecasting wave heights across periods of seven, 14, 30, 45, and 60 days in the Pangandaran area of Indonesia. The research employed five years of training data with four, five, and six months of experimental data. Their investigation demonstrated that XGBoost yielded superior results compared to AdaBoost; XGBoost achieved a predictive evaluation metric RMSE of 0.093 and a correlation coefficient (CC) of 0.989, whereas AdaBoost attained an RMSE of 0.110 and a CC of 0.989. XGBoost and AdaBoost, while widely utilised algorithms possess inherent restrictions. XGBoost is prone to overfitting with noisy data, but AdaBoost is suboptimal for time series data characteristics [5], [6]. Elsworth et al. [7] Also Executed a study with the LSTM model. The LSTM model attains an overall average score of 88.39 across the all-time series, with an average calculation duration of 1293 seconds per time series. The significant calculation time, although precise, is a substantial limitation due to the complexity of optimizing the LSTM model parameters.

This research uses a machine learning called the CatBoost model to forecast wave data and seeks to improve resolution across networks worldwide. CatBoost is intended to provide superior performance and fast training duration to handle massive data sets while maintaining efficiency [8]. Additionally, the model employs strategies to reduce the risk of overfitting, thereby increasing its capacity to generalise previously unseen data [9]. This study uses CatBoost to handle wave height prediction in a region affected by significant ocean waves. In this paper, we examine the coastal area of Pacitan Beach in East Java, Indonesia, which is subject to considerable wave and surge action from the Indian Ocean. We performed high-resolution nested wave simulations utilising the SWAN model to get precise wave data for the Pacitan coastal region[1], [10], [11]. We assessed the efficacy of the CatBoost model in predicting wave heights 1, 5, 7, and 14 days in advance across various situations with differing training data durations over periods of 1, 2, and 8 years. The CatBoost model was optimised for hyperparameters using the GridSearch approach. Furthermore, its efficacy was compared to other prevalent machine learning models for time series forecasting, such as XGBoost and AdaBoost. Root Mean Square Error (RMSE) and R-squared (R^2) measures evaluate the model's efficacy.

This paper's structure is as follows: Section II discusses the literature review on wave height forecasting, CatBoost, XGBoost, and AdaBoost. Section III describes the methodology used. Section IV analyses the forecasting outcomes obtained from XGBoost and AdaBoost. Section V presents the conclusion of this study.