Jurnal Pijar MIPA
https://jurnalfkip.unram.ac.id/index.php/JPM

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

Tidal Flood Prediction in Surabaya Based on Hydrometeorological Data Using Gradient Boosting and Logistic Regression

Kartika Dwi Indra Setyaningrum[1], Kiki Syalasyatun Masfufah[1], Endah Rahmawati[1], Ady Hermanto[2]

[1]Department of Physics, Universitas Negeri Surabaya, Indonesia
[2]BMKG, Tanjung Perak Surabaya, Indonesia
[*]e-mail: endahrahmawati@unesa.ac.id

Received: August 27, 2025. Accepted: October 1, 2025. Published: October 15, 2025

Abstract: This research aims to develop a predictive model for tidal inundation at Tanjung Perak Port in Surabaya, a region identified as critical and highly susceptible to such events. The foundational data incorporated comprises hydrometeorological indicators, such as lunar cycles, tidal patterns, and precipitation levels, which were sourced from the BMKG Tanjung Perak Maritime Meteorological Station. A dataset comprising 26,275 individual data points was compiled and subsequently partitioned into training sets (80% of the data) and validation sets (20%) via randomization. This apportionment is intended to support the robustness and applicability of the developed model. The initial data preparation phase involved techniques such as data normalization, imputation of missing values, and the determination of variable weights based on their respective degrees of impact. Subsequently, two distinct machine learning methodologies were employed to construct the predictive framework: Gradient Boosting (specifically, XGBoost) and Logistic Regression. The efficacy of the resultant models was rigorously assessed using various metrics, including accuracy, confusion matrix analysis, ROC-AUC scores, and feature significance analysis. Analysis of the outcomes indicated that the Gradient Boosting model achieved a superior accuracy of 99.96%, whereas Logistic Regression attained 99.85%. An examination of the features revealed that lunar cycles and tidal conditions were the principal determinants of tidal inundation, with precipitation exerting a comparatively minor effect. These observations substantiate the efficacy of integrating suitable data preparation techniques with machine learning methodologies to achieve precise predictive outcomes. The principal contribution of this investigation is the establishment of a computational framework to facilitate the development of an advanced warning system for tidal flooding, thereby aiding hazard reduction and limiting adverse societal, financial, and operational consequences in littoral regions.

Keywords: Classification; Gradient Boosting; Hydrometeorology; Logistic Regression; Tindal Flooding.

Introduction

Coastal areas play a crucial role in supporting economic activities, particularly in the maritime transportation and interregional distribution sectors. Ports are key hubs in the logistics network, both nationally and internationally, so smooth port operations are crucial to the success of trade and industrial activities [1]. However, coastal areas face major challenges in the form of hydrometeorological disasters, such as tidal flooding. Coastal inundation, characterized by the encroachment of ocean water onto terrestrial areas, is frequently initiated by elevated tidal levels, which are influenced by lunar gravitational forces. Such events are often intensified by severe meteorological conditions, including substantial precipitation. This occurrence has a detrimental effect on various aspects of port operations, notably cargo management. Furthermore, it can lead to the deterioration of civil engineering structures, impede the societal and commercial endeavors of adjacent populations, and engender considerable economic repercussions. Consequently, establishing an accurate system for forecasting tidal flood occurrences is crucial for reducing risk and facilitating adaptation strategies in vulnerable coastal regions. This study was conducted at Tanjung Perak Port in Surabaya, one of Indonesia's major ports, which plays a vital role in national logistics distribution. The high density of industrial and residential activities around the port increases the complexity of the tidal flooding's impacts. This situation requires the application of methods capable of predicting potential flooding with high accuracy so that mitigation measures and policies can be implemented in a timely and effective manner. The primary drivers of tidal inundation events are hydrometeorological conditions, including lunar cycles, tidal elevations, and daily precipitation amounts. Variations in lunar phases induce gravitational stresses that impact oceanic water level fluctuations. Elevated rainfall contributes to increased surface water accumulation, thereby elevating the potential for flooding, particularly when synchronized with peak tidal periods. The intricate interdependencies among these elements produce phenomena that conventional, typically linear or descriptive, forecasting approaches struggle to accurately represent.

Coastal inundation is an intermittent hazard affecting shorelines, occurring when tidal movements combine with various meteorological conditions, causing water to submerge low-elevation zones. Such events present considerable dangers to metropolitan infrastructure, transit

How to Cite:
K. D. I. Setyaningrum, K. S. Masfufah, E. Rahmawati, and A. Hermanto, "Tidal Flood Prediction in Surabaya Based on Hydrometeorological Data Using Gradient Boosting and Logistic Regression", *J. Pijar.MIPA*, vol. 20, no. 6, pp. 1176–1181, Oct. 2025.
https://doi.org/10.29303/jpm.v20i6.10068

systems, and the economic stability of populations residing near the coast. Contemporary investigations have analyzed the synergistic impacts of tides and precipitation on inundation occurrences by employing a unified hydrodynamic methodology and machine learning techniques [2]. This research emphasises the significance of considering nonlinear dynamics and complex relationships among variables in understanding the characteristics of tidal flooding. Nevertheless, a predominant number of current methodologies continue to rely on descriptive examinations or basic statistical techniques, which frequently prove insufficient for adequately depicting the multiple factors that influence tidal inundation.

To overcome these constraints, sophisticated computational techniques have been investigated. Machine learning methodologies, notably, provide a means to represent complex, non-linear associations and interplays with greater proficiency than traditional strategies. The Gradient Boosting approach falls into this category. The Gradient Boosting algorithm, in particular, has demonstrated substantial predictive power. This approach constructs models sequentially by reducing the prediction discrepancies from earlier phases, which consequently facilitates the detection of intricate patterns that are frequently missed by less sophisticated models. [3]. This research employs logistic regression as a comparative methodology alongside Gradient Boosting. Logistic Regression is a well-established statistical technique applied to binary classification tasks. It provides a straightforward and comprehensible approach to analyzing the impact of independent variables on the likelihood of a specific outcome, which in this context pertains to tidal inundation. While Logistic Regression exhibits constraints in modelling nonlinear associations, its strengths include a stable predictive framework and clear interpretability. These attributes render it an appropriate baseline for assessing more sophisticated algorithms. Through the comparison of Gradient Boosting and Logistic Regression performance, this investigation facilitates a thorough evaluation of machine learning techniques for forecasting tidal flooding events at Tanjung Perak Port. The expected contribution is the development of a more reliable early warning system, which can assist policymakers in formulating effective strategies to mitigate hydrometeorological disasters in coastal areas that are increasingly vulnerable to climate and environmental change.

This is reinforced by high-resolution numerical methodologies that consider physical processes in estuarine hydrodynamics, serving as a crucial foundation for obtaining reliable tidal flood predictions and accurately estimating flood extent within estuarine and bay systems [4]. Predictive models based on machine learning, such as Gradient Boosting and Logistic Regression, can utilize these parameters (moon phase, tidal height, rainfall) to more accurately estimate the likelihood of tidal flooding and even estimate flood levels. The integration of data modelling and physical principles is a crucial foundation for developing precise and adaptive early warning systems for hydrometeorological disasters that account for local dynamics [5].

Based on the research results, the Gradient Boosting model achieved the highest accuracy of 99.96%, while Logistic Regression achieved 99.85%, indicating that both models have excellent predictive capabilities, with XGBoost slightly superior.

These results are consistent with the findings of Zhang et al [6], which reported that XGBoost achieved an (Area Under Curve) AUC of 0.94 and an accuracy of 92%, higher than Logistic Regression with an AUC of 0.85. Similar findings were reported by Pratama et al. [7], who found that LightGBM and XGBoost achieved 10–15% higher accuracy than Logistic Regression for flood prediction in Jakarta.

Previous research on the capability of Logistic Regression to identify flood-susceptible areas in a small watershed shows that it is effective for identifying flood-prone areas when the predictor variables are linear, but less optimal for complex spatial data [8]. Conversely, ensemble models such as Gradient Boosting are better able to capture the complex spatial and temporal dynamics in hydrological and estuarine systems. Thus, integrating hydrodynamic physical modelling with a Gradient Boosting-based machine learning approach is a crucial step toward developing a precise, adaptive, and context-aware early warning system for the local dynamics of Tanjung Perak Port.

Although several studies in Indonesia and Southeast Asia have examined tidal flooding using statistical approaches, remote sensing, or deep learning, such as Water Level Time Series Forecasting Using TCN in Surabaya [9], deep Learning for Tidal Flood Prediction in West Pandeglang, Banten [10], and Tidal Flood Vulnerability Assessment in Central Java [11], most of them focus on the influence of a single variable, regional vulnerability, or the use of a single algorithm, and rarely evaluate the importance of variable features in depth. In contrast, this study presents a comparative machine learning framework that compares Gradient Boosting and Logistic Regression, employing comprehensive preprocessing and feature importance analysis to determine the relative contributions of the moon phase, tides, and rainfall to the potential for tidal flooding at Tanjung Perak Port.

Research Methods

This study began with the collection of hydrometeorological data from the Meteorology, Climatology, and Geophysics Agency (BMKG), specifically the Tanjung Perak Surabaya Maritime Meteorological Station. Three main variables were used: moon phase, tidal range, and daily rainfall. The initial data were obtained in the form of graphs or PDF files from routine observations, which were then manually converted into Excel format for quantitative processing.

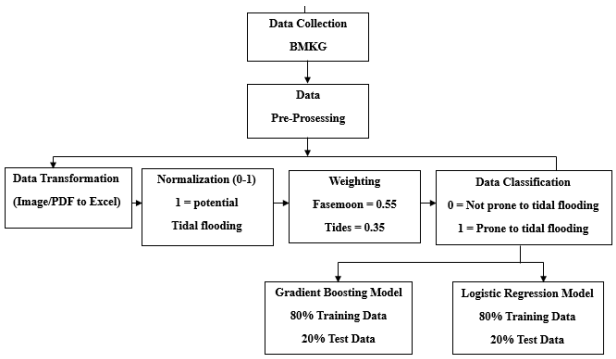


Figure 1 Workflow of research

After the data collection stage was completed, a pre-processing process was carried out consisting of three main steps, namely converting the data from image format to Excel format, normalizing all variables to a scale of 0–1 (where a value of 1 indicates the condition with the highest potential for tidal flooding), and weighting the variables according to their level of contribution to tidal flooding events, namely the moon phase at 0.55, tidal range 0.35, and rainfall 0.10. The objective of the normalization and weighting stages is to prevent any individual variable from disproportionately influencing the model's learning curve and to ensure that each parameter's impact is commensurate with its importance. Contemporary research has frequently used normalization methods, such as min–max scaling and Z-score standardization, to reconcile variables measured in disparate units. Concurrently, weighting approaches such as entropy, AHP, or PCA are applied to assign appropriate weights to each factor that affects flood vulnerability [12].

After normalization and weighting, the collected indicators are integrated to form a composite flood potential index. This consolidation effectively distils complex, multidimensional environmental information into a format that is readily understandable on a numerical scale. Subsequently, this composite index is segmented into distinct classifications, each corresponding to a specific degree of flood risk. For instance, a value of 0 might denote the absence of tidal flooding, while a value of 1 could signify severe tidal flooding. This classification enables more precise spatial identification of areas susceptible to flooding. Contemporary studies underscore that the method selected for normalization, weighting, and classification has a substantial impact on the ultimate cartographic representation and the ease with which flood potential outcomes can be understood. This consequently emphasizes the critical need for both methodological uniformity and thorough sensitivity analysis[13].

Next, the dataset was divided into two parts: 80% for the training set (21,020 data points) and 20% for the test set (5,255 data points). This division aimed to evaluate the model's ability to generalize previously unseen data. In this study, two algorithms—Gradient Boosting and Logistic Regression—were used to model the potential for tidal flooding. The Gradient Boosting technique was selected for its ability to construct models sequentially. This approach refines predictions iteratively by addressing prior errors and is adept at discerning intricate, non-linear relationships within data. Logistic Regression is employed as a benchmark, providing a more straightforward and easily understood model to explain the relationship between predictor variables and the likelihood of tidal flooding. Model efficacy is assessed through accuracy metrics, confusion matrices, and an examination of feature significance to identify the variables that exert the greatest influence on classification outcomes. Recent literature shows that combining ensemble tree methods with general linear models outperforms simpler approaches for flood and hazard prediction, both in terms of precision and interpretability [14]. With this approach, the research not only produces a reliable prediction model but also provides a comprehensive assessment of the model's stability as a supporting component of an early warning system in coastal areas.

Results and Discussion

Hyperparameters of Gradient Boosting and Logistic Regression

To obtain the optimal model configuration, several combinations of hyperparameters were tested, such as max_depth, min_samples_split, and n_estimators. The goal was to find the configuration that produced the highest accuracy in the classification process.

Table 1. Hyperparameters for gradient boosting

Max depth	Min samples split	n_estimators	Accuracy
3	2	200	0.99961941
10	2	50	0.99961941
3	10	100	0.998287345
3	2	50	0.994671741
3	5	50	0.994671741

Various combinations were systematically tested, and the test results showed that the optimal combination was achieved with the settings max_depth = 3, min_samples_split = 2, and n_estimators = 200, yielding a model with an accuracy of 99.96%. This arrangement maintains the model's straightforwardness and prevents excessive fitting to the training data, while simultaneously enabling progressive error refinement via the boosting process.

The visualization presented delineates the outcomes of optimizing the hyperparameters for the Logistic Regression algorithm. This assessment involved a methodical exploration of various settings for the regularisation strength (C) and the optimisation algorithm employed. The C parameter controls the strength of regularization imposed on the model. An elevated C value allows the model to capture intricate data patterns more effectively, while a reduced C value enforces stronger regularization, resulting in a simpler model, though this simplification may compromise predictive accuracy. The regularization techniques, identified as l1 and l2, were implemented to control the regularization process, with C dictating the magnitude of this control (a larger C signifies less regularization). In parallel, optimization routines such as liblinear and lbfgs were utilized for model fitting. Academic discourse has addressed comparable methodological considerations concerning the effect of regularization on logistic regression within high-dimensional environments [15]. This optimization process aims to determine the configuration that yields superior classification precision, while concurrently balancing model complexity and predictive performance across independent datasets.

While Gradient Boosting is recognized for its ability to achieve high accuracy, this modelling approach has certain limitations. A notable disadvantage is its susceptibility to overfitting, especially when used with small or imbalanced datasets. Furthermore, Gradient Boosting requires meticulous tuning of its hyperparameters to achieve peak performance; consequently, selecting unsuitable parameter values can compromise the model's predictive generalization. From a computational standpoint, this methodology is comparatively more demanding than simpler algorithms, resulting in longer model training times and

substantial memory requirements. This research underscores that the problem of overfitting within Gradient Boosting can be addressed through the implementation of cross-validation and thorough data preprocessing methods, thereby improving the model's capacity to generalize to unseen data [16].

Table 2 Logistic Regression Hyperparameters

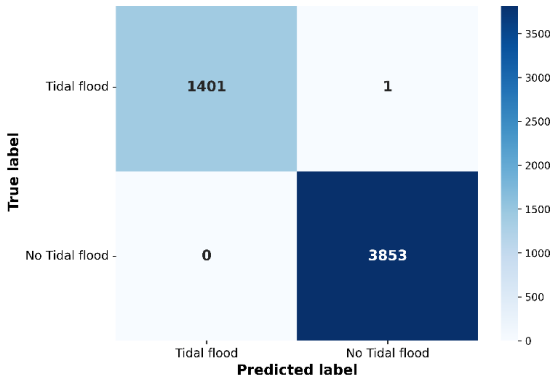
Penalty	C	Solver	Accuracy
l1	2	liblinear	0.99847764
l1	1	liblinear	0.99809705
l1	0,5	liblinear	0.996574691
l2	2	lbfgs	0.971075167
l2	1	lbfgs	0.970313987
l2	0,5	lbfgs	0.967269267

Analysis of the data presented in the table indicates that Model 7, characterized by an L1 configuration penalty, a C value of 2, and the liblinear solver, yielded the optimal outcome. This configuration achieved the highest accuracy rate at 99.85%, as highlighted in yellow. Conversely, Model 2, employing an L2 penalty, a C value of 0.5, and the lbfgs solver, yielded the lowest accuracy of 96.73%. This disparity underscores the substantial influence that the choice of regularization type, regularization strength, and optimization algorithm can exert on a model's overall effectiveness. Notwithstanding these variations, all evaluated configurations demonstrated a commendable accuracy exceeding 96%, implying that the dataset possesses sufficiently discernible patterns to be effectively delineated using Logistic Regression.

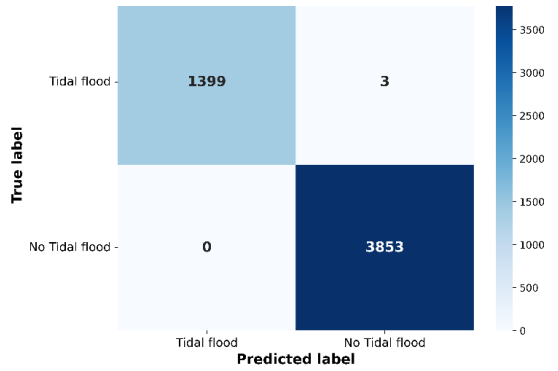
Logistic Regression has several important limitations, despite its frequent use, due to its simplicity and ease of interpretation. One limitation is that when the outcome (dependent variable) is a relatively common event (>10%), the odds ratio estimate from Logistic Regression can be biased when the underlying assumption is prevalence or risk (prevalence ratio/relative risk) [17]. In addition, Logistic Regression tends not to handle highly complex or non-linear data structures well unless feature transformations or variable interactions are explicitly applied. This model can also be disrupted by multicollinearity among predictor variables, and its performance is suboptimal when the data are incomplete or contain many missing values [18].

Confusion Matrix of Gradient Boosting and Logistic Regression

Evaluating model performance using a matrix provides a clear picture of the classification accuracy achieved by each method. Both the Gradient Boosting technique and Logistic Regression exhibit strong predictive performance, with minimal misclassifications observed in the evaluation dataset. Gradient Boosting achieves superior accuracy, whereas Logistic Regression, despite a marginal decrease in accuracy, yields dependable outcomes and provides enhanced clarity regarding the influence of predictor variables.



(a)



(b)

Figure 2. Confusion matrix of (a) Gradient Boosting, (b) Logistic Regression.

The Gradient Boosting model performed excellently when evaluated using a confusion matrix. Analysis of the test results indicates that the model accurately classified 1,401 instances within the "not prone to tidal flooding" group, with a single misidentified case. Conversely, within the "potential tidal flooding" classification, all 3,853 data entries were accurately assessed, resulting in no false negatives. These findings demonstrate Gradient Boosting's proficiency in discerning data trends with high accuracy, encompassing both the identification of secure environments and the recognition of hazardous situations. With a total accuracy of 99.98%, the model has demonstrated its high reliability for integration into early warning frameworks. The low incidence of predictive errors is a crucial benefit, serving to prevent the dissemination of erroneous information, avert public alarm, and ensure that intervention measures are implemented only when demonstrably warranted.

Despite a marginally lower accuracy than Gradient Boosting, the Logistic Regression model demonstrates commendable efficacy. Across the entire test data set, this model correctly predicted 1,399 data points in the “potential tidal flooding” category. There were only three cases that should have been classified in this category but were instead predicted as “not at risk of tidal flooding,” resulting in false negatives. Regarding the classification of areas not susceptible to tidal flooding, Logistic Regression exhibited complete accuracy, correctly identifying all 3,853 instances without any erroneous predictions. The model's overall performance registered an accuracy rate of 99.92%.

A significant advantage of Logistic Regression is its ability to clearly illustrate how specific predictor variables influence the probability of a particular event. This characteristic facilitates the analysis and subsequent

communication of results to relevant parties. Therefore, despite Gradient Boosting's greater predictive power, Logistic Regression remains a viable option, especially when understanding the underlying relationships is of primary importance. For instance, the study titled Gradient Boosting achieves higher accuracy than Logistic Regression with large datasets. Logistic Regression is more readily explainable to individuals without technical expertise due to its directly interpretable coefficients, as shown in the study of Seto et.al [19].

A Comparative Examination of Feature Significance in Gradient Boosting and Logistic Regression Models

The comparative evaluation of feature importance across these two modelling approaches indicates that the lunar phase consistently emerges as the predominant determinant of tidal inundation likelihood. In the Gradient Boosting model, the lunar phase contributes nearly 0.7, whereas in Logistic Regression, it reaches around 30, which is significantly greater than that of other variables. This suggests that astronomical factors, specifically the position and phase of the moon, are closely linked to the dynamics of tidal movements, which are the primary drivers of tidal flooding. Meanwhile, tidal parameters rank second in both models, with significant importance levels of approximately 0.3 in Gradient Boosting and 20 in Logistic Regression, thereby continuing to play a major role in enhancing the predictive model's accuracy.

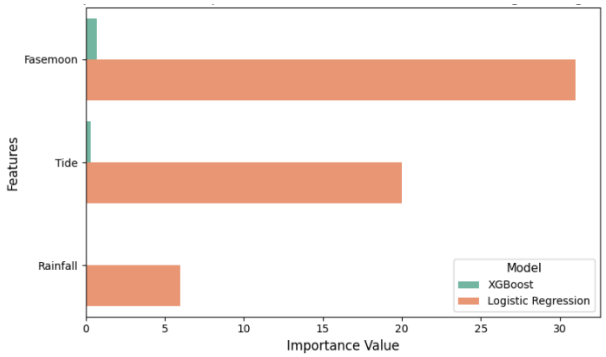


Figure 3 Comparison of the Importance of Features in Gradient Boosting and Logistic Regression

Rainfall in both models showed the lowest contribution, with values of around 0.03 in Gradient Boosting and 5 in Logistic Regression. While its influence is minor compared to celestial phenomena and tidal forces, precipitation remains a contributing factor that can increase the likelihood of tidal inundation when it coincides with peak tides. Consequently, the findings from this investigation substantiate the notion that lunar cycles and tidal patterns warrant primary consideration in tidal flood warning systems, whereas rainfall is better characterized as an auxiliary factor that amplifies the potential for tidal flooding.

Evaluation of XGBoost and Logistic Regression Efficacy Using ROC-AUC

An examination of the preceding ROC Curve illustration reveals that both the XGBoost and Logistic Regression models achieve exceptionally high performance, as evidenced by their respective Area Under the Curve (AUC) values of 0.999987 and 0.999996. An AUC value close to 1 indicates that both models can distinguish between positive and negative classes with near-perfect accuracy. In theory, AUC measures a model's ability to perform correct classification at various thresholds. The higher the AUC value, the better the model's ability to identify the target category.

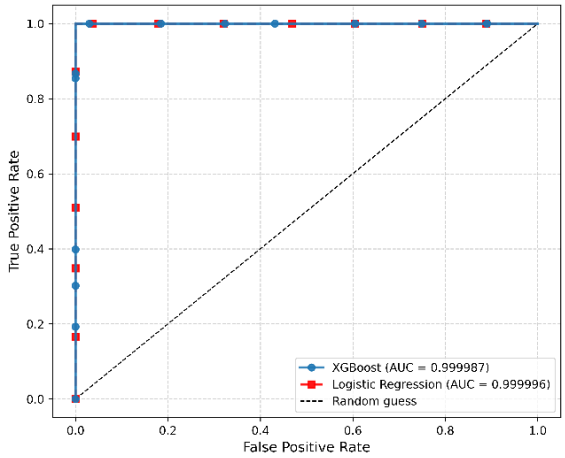


Figure 6. ROC Comparison Graph

The Receiver Operating Characteristic (ROC) curves for both demonstrated models display considerable convergence in the upper-left region. This pattern indicates outstandingly high true positive rates and remarkably low false positive rates. These findings highlight the effectiveness of both Gradient Boosting and Logistic Regression in accurately identifying tidal flooding while simultaneously reducing classification errors. Although the difference is minimal, the Area Under the Curve (AUC) for Logistic Regression shows a slight advantage, suggesting a modest improvement in its generalization. Nevertheless, from a practical perspective, the predictive power of both models is effectively equivalent. Current investigations suggest that Logistic Regression models exhibit performance levels comparable to those of ensemble techniques, including gradient boosting, particularly in areas such as calibration and decision analysis. This is especially true when the dataset is of good quality and lacks excessive complexity [20].

Conclusion

Based on the study's findings, both Gradient Boosting and Logistic Regression demonstrate significant capacity to accurately forecast the likelihood of tidal inundation. The Gradient Boosting model with the optimal hyperparameter configuration (max_depth = 3, min_samples_split = 2, n_estimators = 200) achieved the highest accuracy of 99.96%, demonstrating its ability to recognize complex data patterns. Meanwhile, Logistic Regression achieved an optimal accuracy of 99.85% with a penalty of l1, C = 2, and solver = liblinear, demonstrating excellent performance and the advantage of ease in interpreting the influence of predictor variables. An examination of feature significance revealed that lunar phase was the primary determinant, with tidal influences ranking second, and precipitation serving as a supplementary factor. The proximity of the ROC-AUC values for both models to unity substantiates their efficacy in

classification tasks. Furthermore, these outcomes may offer practical examples for physics and environmental education, illustrating the use of data-driven methodologies to understand natural phenomena. In addition, the results support environmental management efforts, particularly in developing early warning systems and adaptive strategies to mitigate the impacts of tidal flooding on coastal communities.

Author's Contribution

Kartika Dwi Indra Setyaningrum: was responsible for the conceptualization of the research, data collection, and data analysis; Kiki Syalasyatun Masfufah: contributed to data processing; Endah Rahmawati: supervised the research, methodological review, and validation of the analysis results; Ady Hermanto: provided data and supervised data collection.

Acknowledgements

The author would like to express to the Meteorology, Climatology, and Geophysics Agency (BMKG) of Tanjung Perak Surabaya for providing the data used in this study.

References

[1] N. Grubišić, T. Krljan, and K. Sesar, "Traffic microsimulation of the main junction connecting the urban road network with the sea-port container terminal," *Pomorstvo*, vol. 37, no. 1, pp. 106–117, 2023, doi: 10.31217/p.37.1.9.

[2] J. Sampurno, V. Vallaeys, R. Ardianto, and E. Hanert, "Integrated hydrodynamic and machine learning models for compound flooding prediction in a data-scarce estuarine delta," *Nonlinear Process. Geophys.*, vol. 29, no. 3, pp. 301–315, 2022, doi: 10.5194/npg-29-301-2022.

[3] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[4] C. L. Lopes *et al.*, "Evaluation of future estuarine floods in a sea level rise context," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-12122-7.

[5] H. Daher *et al.*, "Long-Term Earth-Moon Evolution With High-Level Orbit and Ocean Tide Models," *J. Geophys. Res. Planets*, vol. 126, no. 12, Dec. 2021, doi: 10.1029/2021JE006875.

[6] Y. Wu, Z. Zhang, X. Qi, W. Hu, and S. Si, "Prediction of flood sensitivity based on Logistic Regression, eXtreme Gradient Boosting, and Random Forest modeling methods," *Water Sci. Technol.*, vol. 89, no. 10, pp. 2605–2624, 2024, doi: 10.2166/wst.2024.146.

[7] C. Gde and L. Pringandana, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," vol. 6, no. 4, pp. 2467–2483, 2025.

[8] M. L. Edamo, E. G. Ayele, T. Y. Ukumo, A. A. Kassaye, and A. P. Haile, "Capability of logistic regression in identifying flood-susceptible areas in a small watershed," *H2Open J.*, vol. 7, no. 5, pp. 351–374, 2024, doi: 10.2166/h2oj.2024.024.

[9] D. Saepudin, E. S. Rabbani, D. Navialdy, and D. Adytia, "Water Level Rise Forecasting Using TCN Study Case in Surabaya Coastal Area," *J. Online Inform.*, vol. 9, no. 1, pp. 61–69, 2024, doi: 10.15575/join.v9i1.1312.

[10] W. P. Waters, N. A. Ramaputra, A. S. Budiman, and W. A. Arifin, "Deep Learning for Tidal Flood cPrediction in," vol. 10, no. 1, 2025.

[11] K. Adillah, A. Sakti, L. Syahid, and K. Wikantika, "Assessing Tidal Flooding Vulnerability in the Coastal Region of Central Java Using Remote Sensing Approach," 2024, doi: 10.4108/eai.24-11-2023.2346418.

[12] L. L. Moreira, M. M. de Brito, and M. Kobiyama, "Effects of different normalization, aggregation, and classification methods on the construction of flood vulnerability indexes," *Water (Switzerland)*, vol. 13, no. 1, 2021, doi: 10.3390/w13010098.

[13] B. Ma *et al.*, "Comprehensive risk assessment of urban floods based on flood simulation and socio-economic vulnerability," *Front. Earth Sci.*, vol. 13, no. August, pp. 1–16, 2025, doi: 10.3389/feart.2025.1645693.

[14] J. Zhang, W. Guo, S. W. Chang, D. D. Nguyen, and H. H. Ngo, "Data-Driven Innovations in Flood Hazard Assessment with Machine Learning," 2025.

[15] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1–25, 2019.

[16] Q. A. Hidayaturrohman and E. Hanada, "Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure," *BioMedInformatics*, vol. 4, no. 4, pp. 2201–2212, 2024, doi: 10.3390/biomedinformatics4040118.

[17] L. Pinheiro-Guedes, C. Martinho, and M. R. O. Martins, "Logistic Regression: Limitations in the Estimation of Measures of Association with Binary Health Outcomes," *Acta Med. Port.*, vol. 37, no. 10, pp. 697–705, 2024, doi: 10.20344/amp.21435.

[18] N. R. Panda, J. K. Pati, J. N. Mohanty, and R. Bhuyan, "A Review on Logistic Regression in Medical Research," *Natl. J. Community Med.*, vol. 13, no. 4, pp. 265–270, 2022, doi: 10.55489/njcm.134202222.

[19] H. Seto *et al.*, "Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022, doi: 10.1038/s41598-022-20149-z.

[20] A. F. Militino, H. Goyena, U. Pérez-Goya, and M. D. Ugarte, "Logistic regression versus XGBoost for detecting burned areas using satellite images," *Environ. Ecol. Stat.*, vol. 31, no. 1, pp. 57–77, 2024, doi: 10.1007/s10651-023-00590-7.