

MLSEL: A tuned multilayer stacking ensemble learning with meta feature for flood risk prediction

Rini Sovia^{*1)}, Yuda Irawan²⁾, Irzal Arief Wisky¹⁾, Muhammad Habib Yuhandri¹⁾ and Randy Permana¹⁾

¹⁾Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang 25147, Indonesia

²⁾Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru 28288, Indonesia

Received 14 July 2025
Revised 20 November 2025
Accepted 8 January 2026

Abstract

While stacking ensemble methods have been widely adopted for disaster risk classification, most traditional implementations are limited to shallow single-layer architectures, lacking generalization capacity and adaptive meta-feature design. This study proposes Multilayer Stacking Ensemble Learning (MLSEL), a novel approach that addresses these limitations through three key innovations: (i) a deep three-layer stacking architecture, (ii) Meta Feature Augmentation (MFA) to enrich inter-layer representations, and (iii) automated hyperparameter optimization using Optuna to enhance meta-learner performance. The model is evaluated on a multiclass flood risk dataset consisting of 50,000 structured records. Results reveal that the proposed MLSEL particularly the configuration using XGBoost + Optuna + MFA achieves superior accuracy of 98.69%, significantly outperforming both the best-performing baseline and conventional stacking models. This research demonstrates that combining deep-layered learning, meta-feature engineering, and adaptive optimization effectively overcomes overfitting, feature redundancy, and scalability issues inherent in traditional stacking. The MLSEL framework establishes a robust foundation for accurate and reliable disaster risk prediction systems.

Keywords: Multilayer stacking, Stacking ensemble learning, Meta feature, Flood risk prediction, Classification

1. Introduction

Floods remain one of the most pervasive and devastating natural disasters globally, with significant socioeconomic and environmental consequences. Rapid urbanization, climate variability, inadequate infrastructure, and poor watershed management continue to exacerbate the frequency and severity of flooding, particularly in developing regions [1-4]. Consequently, there is a growing demand for robust flood risk prediction systems that are not only accurate but also adaptive to complex data environments [5, 6]. In recent years, the application of machine learning (ML) has shown promise in enhancing flood risk classification and early warning systems due to its ability to capture non-linear relationships in multivariate data [7-12].

Despite the progress, conventional ML models such as Support Vector Machines (SVM), Random Forests (RF), K-Nearest Neighbors (KNN), Extra Trees, and Gradient Boosting Machines (GBM) often rely on singular representations of features and lack the structural depth required for modeling high-dimensional interactions [13-17]. To address this limitation, ensemble learning techniques particularly stacking ensemble learning (SEL) have gained traction [18-20]. These methods aggregate predictions from diverse base learners through a meta-learner, aiming to improve generalization and reduce individual model bias [21-23]. However, most existing SEL implementations are restricted to shallow, single-layer architectures, which can lead to redundant feature learning and suboptimal performance in multiclass settings [24]. The absence of automated hyperparameter tuning and meta-feature design often hinders model robustness and scalability.

Several previous studies have successfully applied the stacking ensemble approach in various domains to improve prediction accuracy. Researchers developed stacking models based on LightGBM, RF, and BPNN with a Linear Regression meta-learner to predict heat and mass transfer in desiccant wheels, accompanied by Bayesian optimization and interpretability using SHAP [25]. Meanwhile, other researchers proposed stacking visual-fault classification for photovoltaic modules with a feature approach from AlexNet and J48 feature selection, followed by testing a combination of base-learner and meta-classifier [26]. Another study applied stacking ensemble to glacier classification using Landsat satellite imagery and recorded an accuracy of up to 94.09% [27]. Although these three studies demonstrated that stacking can improve prediction performance, the approaches used were generally limited to a single layer with a static ensemble architecture and without exploration of feature augmentation or further optimization of the meta-structure.

To fill these gaps, this study proposes a novel Multilayer Stacking Ensemble Learning (MLSEL) model that combines three key innovations: (i) a hierarchical three-layer stacking architecture with diverse learners at each layer, (ii) meta feature augmentation (MFA) to enrich feature space and mitigate information loss across layers, and (iii) Optuna-based hyperparameter tuning at the meta level to optimize model performance. The proposed MLSEL is tested on a flood risk dataset comprising 50,000 structured records with multiclass labels, and further benchmarked against five state-of-the-art SEL frameworks from prior studies.

*Corresponding author.

Email address: rini_sovia@upiyptk.ac.id

<https://doi.org/10.64960/easr.2026.263071>

This study is contextually motivated by flood prone low lying catchments in Southeast Asia, where monsoon driven rainfall patterns, rapid urbanization, wetland degradation, and inadequate river management frequently contribute to severe flooding. The variables incorporated in our dataset such as monsoon intensity, river management effectiveness, urbanization rate, drainage system capacity, deforestation, and wetland loss closely reflect the environmental conditions commonly observed in these regions. Although the dataset used in this research is structured and generalized, it is designed to emulate the hydrological and infrastructural characteristics typical of Southeast Asian monsoon affected floodplains, ensuring the relevance and applicability of the proposed MLSEL model to real world flood prediction scenarios.

The primary contributions of this work are as follows: First, we introduce a deep ensemble architecture for structured tabular flood data, leveraging hierarchical stacking to incrementally capture complex interactions. Second, we incorporate MFA to preserve and enhance feature diversity at the meta level, a component largely overlooked in existing literature. Third, we apply Optuna optimization to refine meta-learner performance through systematic hyperparameter exploration. Finally, we demonstrate the generalizability of the model across datasets and domains, underscoring its potential for broader disaster risk classification applications.

Given these contributions, the proposed MLSEL framework not only addresses key limitations in existing ensemble-based flood classification models but also introduces a scalable architecture suitable for real-world implementations. By integrating layer-wise learning, feature enhancement, and automated tuning, this research sets a new benchmark for intelligent disaster prediction systems. This work is anticipated to inform future studies in ensemble learning design, particularly in multiclass risk assessment domains, and to provide decision-makers with a more reliable basis for proactive flood mitigation planning.

2. Materials and methods

Stacking ensemble is a popular approach in machine learning that aims to improve prediction accuracy by combining the outputs of multiple base models [28]. This strategy works by training a meta-classifier to learn how to best combine predictions from previous models [29]. The following Figure 1 presents the basic architecture of the traditional ensemble stacking method.

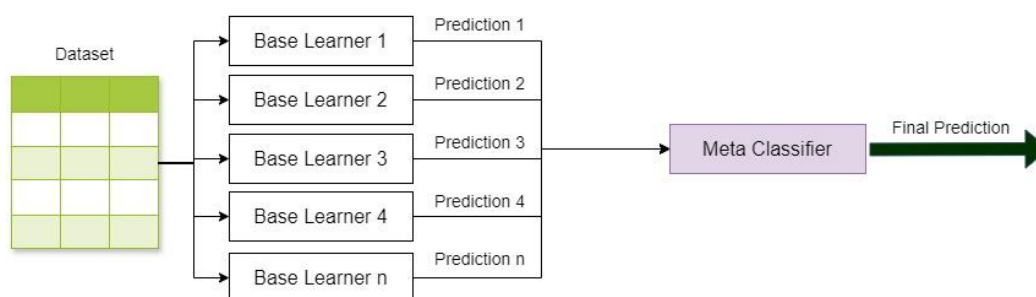


Figure 1 Basic Architecture of the Stacking Ensemble Model

Figure 1 shows the conventional ensemble stacking structure, where multiple base learners are trained in parallel using the same dataset, and their prediction results are used as input features for the meta-classifier [24]. This approach allows the system to integrate the advantages of various base algorithms to improve generalization ability [30]. However, this model has several limitations, including the lack of feature interaction between layers, potential redundancy if the base learner is too uniform, and the risk of overfitting in the meta-classifier if not balanced with cross-validation techniques or feature augmentation [21]. Therefore, the development of a more adaptive and optimized stacking model is needed to produce more reliable predictions. The proposed stacking model consists of three nested layers to accommodate the gradual and systematic prediction aggregation process; the proposed multilayer architecture can be seen in the following Figure 2.

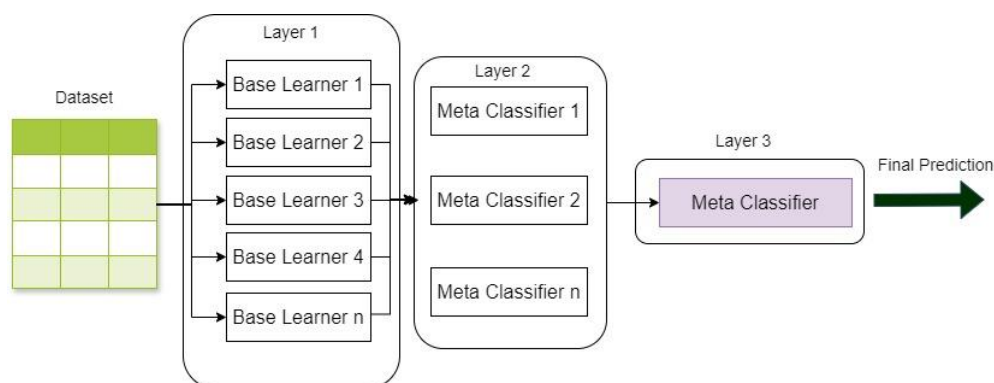


Figure 2 Proposed Multilayer Stacking Ensemble Architecture

Figure 2 shows the structure of a three-layer stacking ensemble, where the learning process starts with several base learners at Layer 1 that perform initial training on the input data [24]. The predictions from each base learner are then passed to Layer 2, which consists of a number of meta-classifiers, which combine the predictions in a more complex and layered manner. The final result from Layer 2 becomes the input for Layer 3 as the final meta-classifier to produce the final prediction. The advantage of this approach lies in its ability to model non-linear relationships between previous predictions, increase model representation capacity, and reduce the

risk of overfitting by gradually distributing complexity between layers [24]. This multilevel structure also allows flexibility in the selection of algorithms at each layer, making it adaptive to different types of data and class distribution patterns.

This research proposes a multi-layer stacking model with feature augmentation and hyperparameter tuning approaches. This model is designed to improve the accuracy and generalization of multiclass classification predictions. The development of the proposed new multilayer stacking model can be seen in the following Figure 3.

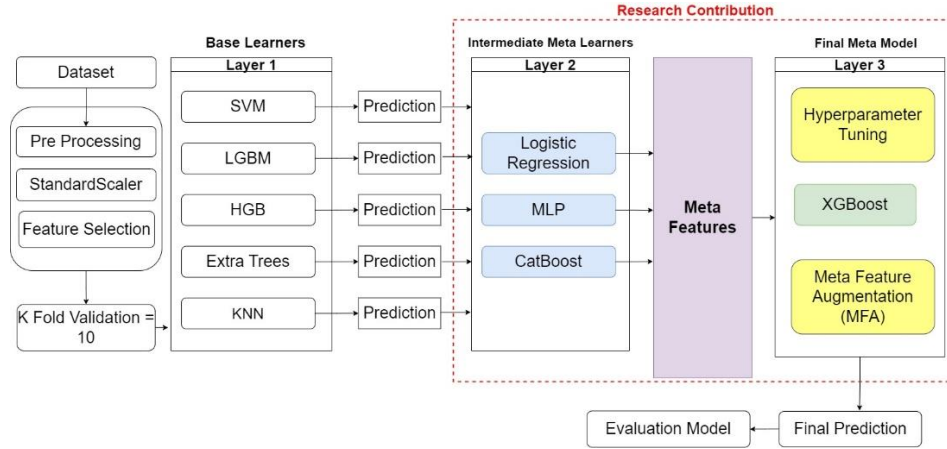


Figure 3 Proposed MLSEL Model

Figure 3 illustrates the innovative architecture of MLSEL, which integrates three learning layers. The first layer consists of several base learners including Support Vector Machine (SVM), Light Gradient Boosting Machine (LGBM), Histogram Gradient Boosting (HGB), Extra Trees Classifier, and K Nearest Neighbors (KNN), which collectively capture diverse representations from the input data. Predictions from this layer become inputs for the second layer which contains intermediate meta learners (Logistic Regression, Multilayer Perceptron (MLP), CatBoost) to perform initial aggregation of detected patterns. The results are converted into a new set of features called meta features and become the input for the final meta model in the third layer. In this layer, meta feature augmentation is performed to enrich the feature representation, followed by hyperparameter tuning using Optuna on the XGBoost model. The main advantage of this architecture lies in its ability to improve generalization ability and diversity between models, while reducing the risk of overfitting through multilevel learning [23]. The innovative combination of meta feature augmentation and hyperparameter optimization makes this architecture a significant novel contribution in the domain of structured data-based multiclass classification [31-34].

The third layer in the MLSEL architecture acts as a final refinement stage that learns higher order interactions between the aggregated meta features. In our comparative experiments, the best two layer configuration which already included hyperparameter tuned XGBoost achieved 98.12 percent accuracy. When the third layer was added, the performance increased to 98.69 percent with reduced variance across folds. This indicates that the additional depth enables the model to capture nonlinear relationships between intermediate level predictions that are not fully learned in a two layer structure. The third layer also stabilizes the ensemble by reducing prediction fluctuations and improving the separation between risk classes.

2.1 Multilayer stacking

The three-layer stacking model developed in this study consists of Layer-1 (base learners), Layer-2 (intermediate learners), and Layer-3 (meta learner). Equations 1 to 10 presented in this section are the original derivations proposed in this study. They formalize the multilayer stacking process, meta-feature generation, and integration of hyperparameter optimization within the MLSEL architecture. Although inspired by the general principles of ensemble learning, the mathematical formulation is not directly adapted from previous publications. Related literature supporting the theoretical foundation of the ensemble stacking method includes previous works [21, 23, 24]. The stacking process is organized as follows:

2.1.1 Layer 1 – base learners

Suppose $B_1, B_2, \dots, B_n$ are base classifiers. For each base model B_i , training was carried out using k-fold cross-validation and produced probabilistic predictions.

$$\hat{Y}_i^{(1)} \in R^{N \times C} \quad (1)$$

Where N is the number of samples and C is the number of classes.

$$X^{(2)} = [\hat{Y}_1^{(1)} \hat{Y}_2^{(1)} \dots \hat{Y}_n^{(1)}] \quad (2)$$

2.1.2 Layer 2 – intermediate learners

The second layer receives the probabilistic output from layer 1 as input. Suppose $L_1, L_2, \dots, L_m$ are intermediate models, each producing a probabilistic prediction:

$$\hat{Y}_j^{(2)} = L_j(X^{(2)}) \text{ for } j = 1, \dots, m \quad (3)$$

The combined predictions of the intermediate layers produce the input for layer 3:

$$X^{(3)} = [\hat{Y}_1^{(2)} \hat{Y}_2^{(2)} \dots \hat{Y}_m^{(2)}] \quad (4)$$

2.1.3 Meta feature augmentation

To enrich the information in the meta-learner, meta feature augmentation is carried out by combining $X^{(3)}$ and the original standardized X_{original} which has been standardized:

$$X^{\text{meta}} = [X^{(3)} X_{\text{original}}] \quad (5)$$

2.1.4 Layer 3 – meta learner

Meta-learner M uses the XGBoost algorithm tuned using Optuna. The goal is to maximize classification accuracy on validation data by finding the optimal hyperparameter combination:

$$\hat{Y}^{\text{final}} = M(X^{\text{meta}}; \theta^*) \quad (6)$$

Where θ^* is the optimal hyperparameter vector resulting from Optuna exploration.

2.2 Data transformation

In the data preprocessing stage, numerical feature transformation is performed using the StandardScaler approach. This method works by calculating the mean (μ) and standard deviation (σ) of each feature, then transforming each X value into a new value X_{standard} , using the formula:

$$X_{\text{standard}} = \frac{X - \mu}{\sigma} \quad (7)$$

This transformation aims to make all features have a mean of zero and standard deviation of one, so that the features are on a comparable scale. This is especially important for machine learning algorithms that are sensitive to data scale, such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). With this standardization, the model training process becomes more stable and accurate as no feature dominates the learning process due to larger scale.

2.3 Meta feature augmentation

In the multilayer stacking ensemble approach, Meta Feature Augmentation (MFA) is a technique that enriches the input feature representation for the meta-learner by combining probabilistic predictions from previous models (base and intermediate learners) with the original features of the dataset. The formulation of MFA is written as follows:

$$Z_{\text{aug}} = [P_{L_k}, X] \quad (8)$$

where: $X \in R^{n \times d}$ is the original feature matrix, consisting of n samples and d features,

$P_{L_k} \in R^{n \times m}$ is the probability matrix of the predicted results from Layer k stacking, where m is the number of class outputs multiplied by the number of models in that layer,

$Z_{\text{aug}} \in R^{n \times (d+m)}$ is the augmented feature that becomes input for the meta-learner in the final layer.

The operation $(,)$ denotes a horizontal concatenation between two feature matrices. In this implementation, P_{L_k} is the output of intermediate learners (Layer-2) consisting of Logistic Regression, MLP, and CatBoost, while X is the original features that have been standardized using StandardScaler.

Beyond the concatenation formulation in Equation 8, the MFA mechanism enriches the representation by generating an intermediate probability space P_{L_k} through the Layer k models. Each probability vector encodes the confidence distribution of the base and intermediate learners across all classes, which introduces higher order predictive signals that are not present in the original feature matrix X . These signals reflect learned nonlinear interactions between features and model specific error patterns.

Algorithmically, the MFA process consists of three main steps. First, each model in Layer k is trained using the outputs from the previous layer. Second, the predicted probability vectors for all samples are collected and stacked to form P_{L_k} . Third, the augmented matrix Z_{aug} is created by concatenating P_{L_k} with X , producing a richer multi scale representation. This augmented representation enables the final meta learner to operate on both physical features and learned predictive patterns.

MFA also reduces redundancy because P_{L_k} represents compact summaries of interactions rather than copies of the original features. Overfitting is mitigated since noisy signals are filtered through multiple learning layers before reaching the final stage. The augmented representation therefore provides stronger generalization capability for the final meta learner.

Conceptually, Meta Feature Augmentation improves model performance through two main contributions. First, MFA expands the meta-learner feature space with probabilistic distribution information captured by previous models, allowing the final model (XGBoost) to recognize predictive patterns that are not detected by a single model. Second, by reinserting the original features into the final input, MFA helps preserve traces of primary information that can be lost due to the layered stacking process, thus supporting better generalization ability to new data.

2.4 Hyperparameter tuning OPTUNA

In the development of MLSEL-based flood risk prediction systems, the selection and tuning of hyperparameters in the meta-learner play a crucial role in improving the accuracy and generalization of the model. In this research, Optuna is used, which is an efficient and automatic Bayesian optimization-based hyperparameter optimization framework to optimize the performance of the XGBoost model as a meta-learner. The optimization process is performed after the formation of meta feature augmentation so that the parameter search process is performed on the final feature representation used by the meta model.

Optuna performs adaptive hyperparameter search using the Tree-structured Parzen Estimator (TPE) approach, which is a probabilistic method for modeling the posterior distribution of an objective function. The objective function $f(\theta)$ is defined as a function that returns an evaluation metric (e.g., accuracy) based on a combination of parameters θ . The goal of this process is to maximize the accuracy function:

$$\theta^* = \arg \max_{\theta \in \Theta} f(\theta) \quad (9)$$

Where, θ is a hyperparameter vector consisting of : n_estimators, max_depth, learning_rate, subsample, colsample_bytree, gamma, reg_alpha, reg_lambda.

$$f(\theta) = \text{accuracy_score}(y_{\text{true}}, \hat{y}) \quad (10)$$

$\hat{y}$ is the prediction result of the XGBoost model on the meta feature set Z_{aug} .

In the implementation, Optuna searches through several iterations (trials) to select the parameter combination that gives the highest accuracy against the validation data. Each iteration will adjust the sampling distribution based on previous results. This process enables a more adaptive and efficient exploration-exploitation approach.

3. Results

3.1 Dataset and preprocessing

The dataset used in this study contains 50,000 entries and 22 features because a large sample size is required to capture the wide variation in hydrological, environmental, and infrastructure conditions that influence flood risk in real settings. Previous studies in flood prediction and environmental modeling have emphasized that models trained on large and heterogeneous datasets are able to represent complex interactions among multiple factors and produce more reliable classification outcomes. The choice of 22 features covers key dimensions commonly associated with flood risk conditions such as river management effectiveness, monsoon intensity, drainage capacity, land use change, wetland degradation, urban expansion, topographic characteristics, and ecosystem pressures.

The hydrological features include monsoon intensity, river management condition, and water flow variability. The environmental features capture land use change, deforestation rate, wetland loss, soil moisture characteristics, and topographic attributes such as local elevation influence and terrain drainage behavior. The infrastructure related features represent drainage system capacity, deteriorating infrastructure, road density, and the extent of urban expansion. Collectively, these features provide a comprehensive representation of the physical and human driven factors contributing to flood risk in monsoon influenced regions. With this number of samples and variables the MLSEL framework can be evaluated robustly and can generalize well across different combinations of risk conditions. The target variable used is the id Flood Risk Level, which represents the level of flood risk and is classified into four classes, namely class 0 (very low), class 1 (low), class 2 (medium), and class 3 (high). The following is Figure 4 regarding class distribution.

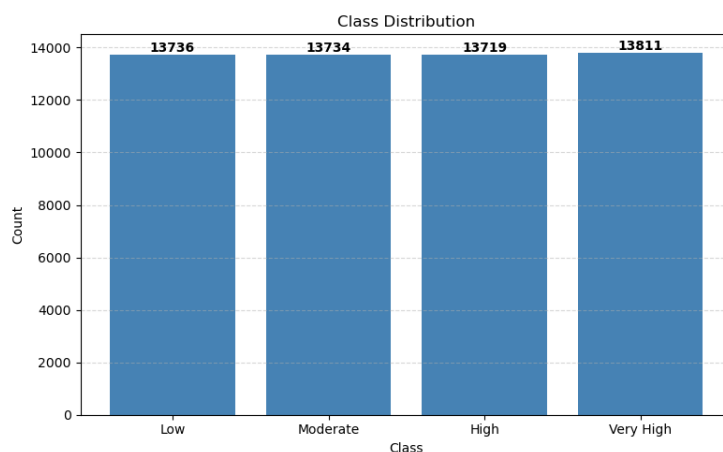


Figure 4 Class Distribution

The distribution of data between classes is quite balanced, with each class having almost the same amount of data: class 0 has 13,736, class 1 has 13,734, class 2 has 13,719, and class 3 has 13,811 entries. This class composition and balance makes the dataset ideal for the development and evaluation of robust multiclass classification models. This balanced structure allows the model to learn all risk levels fairly without domination by a majority class. To understand the contribution of each feature in the classification of flood risk levels, a model interpretability analysis was conducted using the SHAP (SHapley Additive exPlanations) approach. The visualization of the average impact of each feature on the model output can be seen in the following Figure 5.

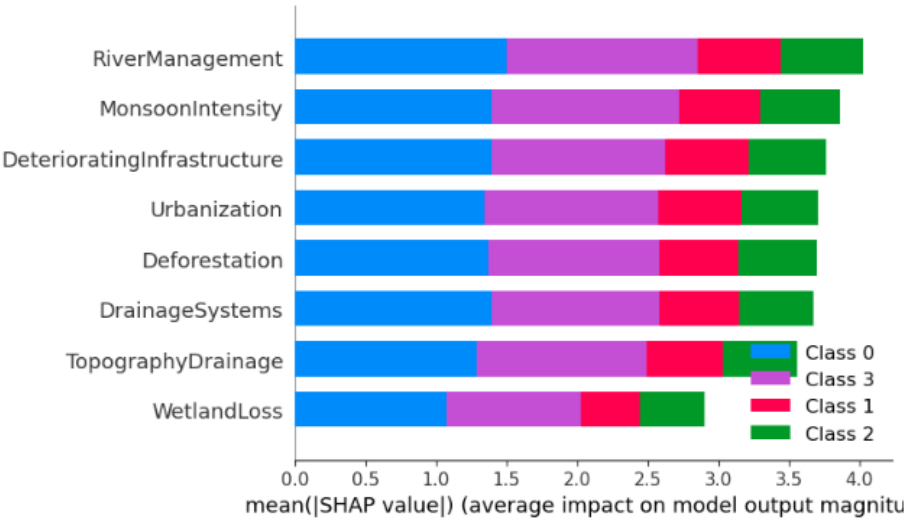


Figure 5 SHAP Summary Bar Chart Visualization of Feature Contribution

Figure 5 shows that the RiverManagement, MonsoonIntensity, and DeterioratingInfrastructure features have the most significant impact on model predictions, especially in classes 0 and 3. This conclusion is based on the SHAP importance ranking, where these three variables have the highest mean absolute SHAP values among all features. Their SHAP distributions show more extreme positive and negative contributions in classes 0 and 3, indicating a stronger influence on the classification of very low and very high flood risk conditions. To gain a visual understanding of the relationships between features and their relevance to the classification target, a correlation matrix visualization was performed in the form of a heatmap. The correlation map of features to flood risk classes can be seen in the following Figure 6.



Figure 6 Heatmap of Correlation Between Features and Flood Risk Classes

Figure 6 shows that most of the features have low correlation with each other, indicated by values close to zero in the area outside the main diagonal, which indicates low multicollinearity between features. The highest correlation to the label is found in the RiverManagement and MonsoonIntensity features with a value of 0.35, while the WetlandLoss feature has the lowest correlation value of 0.26. This pattern suggests that the features used are relatively independent and individually relevant to the target variable, supporting a more stable model learning process and reducing the risk of overfitting. To strengthen the understanding of the relationship between each feature and the flood risk class label, a Pearson correlation analysis was conducted. The correlation value of each feature against the classification target can be seen in the following Table 1.

Table 1 Correlation Value of Features to Flood Risk Levels

Feature	Correlation
RiverManagement	0.354043
MonsoonIntensity	0.353621
DeterioratingInfrastructure	0.334616
Urbanization	0.329581
DrainageSystems	0.328654
Deforestation	0.327759
TopographyDrainage	0.319585
WetlandLoss	0.260781

Table 1 shows that the RiverManagement and MonsoonIntensity features have the highest correlation to flood risk levels, at 0.354 and 0.354 respectively. This indicates that river management effectiveness and monsoon intensity have a relatively strong linear relationship to the risk classification. In contrast, WetlandLoss has the lowest correlation (0.260), but remains relevant in its contribution. This finding is in line with previous SHAP analysis results and provides statistical validation of the importance of these features in the model learning process. Feature selection using the LASSO method was applied to all 22 original features, and the regularization process reduced this set to 8 variables with non zero coefficients. These 8 features were retained because they contributed most strongly to the flood risk classification task.

3.2 MLSEL performance

To assess the effectiveness of conventional stacking methods in multi-class flood risk classification, experiments were conducted using a combination of base learners and meta-models. The performance evaluation includes accuracy, precision, recall, and F1-score metrics to provide an overall picture of the stability and generalization power of the model. The performance results of regular stacking can be seen in Table 2.

Table 2 Performance Evaluation of Baseline and Single-layer Stacking Models

Model	Accuracy (%)	Precision	Recall	F1-Score
SVM	97.68	0.98	0.98	0.98
LGBM	87.43	0.88	0.87	0.87
HGB	88.62	0.89	0.89	0.89
ET	82.80	0.83	0.83	0.83
KNN	82.36	0.83	0.82	0.83
MetaXGB	94.06	0.95	0.94	0.94
MetaXGB + Optuna	95.16	0.96	0.95	0.95
XGB + Optuna + MFA	96.88	0.97	0.97	0.97

Table 2 shows that the SVM model as a single base learner recorded the highest accuracy of 97.68% with consistent precision, recall, and F1-score values of 0.98. However, the conventional stacking approach using XGBoost as a meta-learner showed progressive performance improvement. MetaXGB recorded 94.06% accuracy with an F1-score value of 0.94, which then increased to 95.16% after hyperparameter tuning using Optuna. This increase indicates that parameter optimization automatically contributes to the improvement of model generalization. The integration of Meta Feature Augmentation (MFA) technique in the XGB + Optuna + MFA model resulted in the highest accuracy in the regular stacking group at 96.88% with uniform precision, recall, and F1-score at 0.97. These results show that the combination of optimization strategies and meta-feature augmentation significantly improves the representation of information between model layers, while emphasizing the importance of the architecture refinement process in layered stacking approaches. However, the final accuracy of the stacked model is still unable to surpass the performance of individual SVMs as the best baseline.

The performance evaluation of the MLSEL model is conducted thoroughly on each layer starting from the base learners (Layer 1), intermediate meta learners (Layer 2), to the final meta model (Layer 3). This test aims to observe the contribution of each component to the classification accuracy and consistency of other metrics. The complete test results at all stages of MLSEL can be seen in Table 3.

Table 3 Performance Evaluation of MLSEL Components across Multilayer Architecture

Model	Accuracy (%)	Precision	Recall	F1-Score
SVM	97.68	0.98	0.98	0.98
LGBM	87.43	0.88	0.87	0.87
HGB	88.62	0.89	0.89	0.89
ET	82.80	0.83	0.83	0.83
KNN	82.36	0.83	0.82	0.83
LR (Layer 2)	97.86	0.98	0.98	0.98
MLP (Layer 2)	97.88	0.98	0.98	0.98
CAT (Layer 2)	98.12	0.98	0.98	0.98
MetaXGB (Layer 3)	98.00	0.98	0.98	0.98
MetaXGB + Optuna (Layer 3)	98.36	0.98	0.98	0.98
XGB + Optuna + MFA (Layer 3)	98.69	0.99	0.99	0.99

Table 3 shows that at Layer 1, SVM performed as the best base learner with 97.68% accuracy, far outperforming other models such as LGBM (87.43%) and KNN (82.36%). The performance improvement started to show in Layer 2, where CatBoost obtained 98.12% accuracy with precision, recall, and F1-score of 0.98 consistently, surpassing Logistic Regression (97.86%) and MLP (97.88%). A significant jump occurred at Layer 3 through the use of optimized stacking. The base MetaXGB recorded 98.00% accuracy, increasing to 98.36% when combined with Optuna. The integration of the MFA strategy resulted in the highest performance in the XGB + Optuna + MFA model with an accuracy of 98.69% and another metric score of 0.99. This proves that multilayer architectures and meta-feature augmentation techniques synergistically enrich the representation of data between layers, resulting in models that are more robust and superior to conventional single-layer or stacking approaches.

The visualization of each models performance at various MLSEL layers reinforces the previous numerical findings and provides a clearer comparative perspective of the models generalization capabilities. This graph displays a comparison of the main metrics namely Accuracy, Precision, Recall, and F1-Score of all models tested, both at the base learner and meta learner levels. The complete visualization display can be seen in Figure 7.

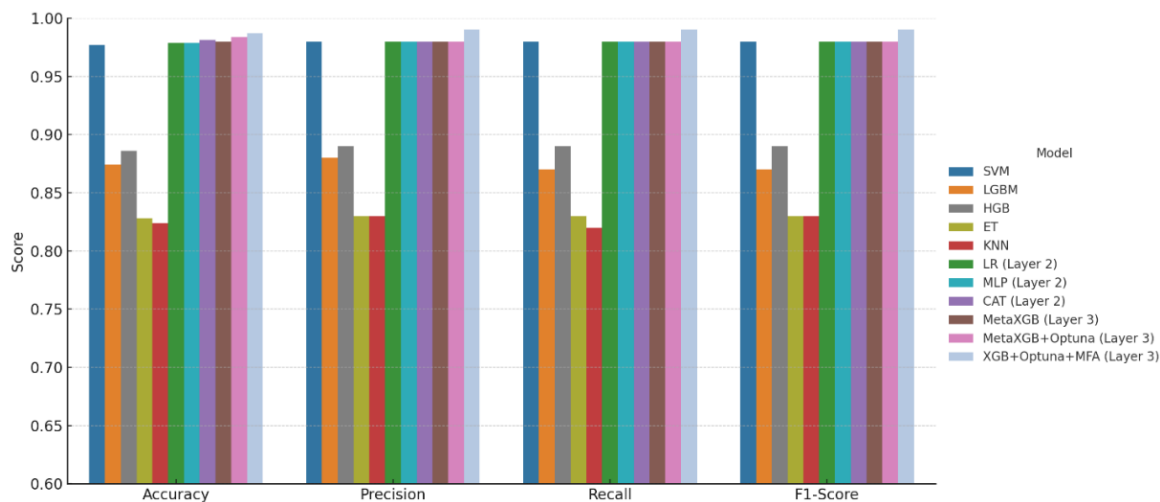


Figure 7 Comparative Performance Metrics of MLSEL Components across Multiple Layers

The graph in Figure 7 shows a consistent performance improvement from base learners to higher meta layers. SVM, the initial base model, scored high across all metrics, at around 0.98. Significant improvement was seen in Layer 2, with CatBoost slightly outperforming SVM. The best performance was achieved in Layer 3, with the XGB+Optuna+MFA model approaching a perfect score of 0.99. These results confirm that the integration of Optuna and MFA in the final layer significantly contributes to the overall accuracy and stability of the model. To evaluate the stability and consistency of model performance during the meta learning stage, a boxplot visualization of the accuracy distribution was performed. This analysis aimed to identify the extent of variation in the performance of each model and the potential improvements resulting from feature optimization and augmentation. The results of the visualization of the accuracy distribution are shown in Figure 8.

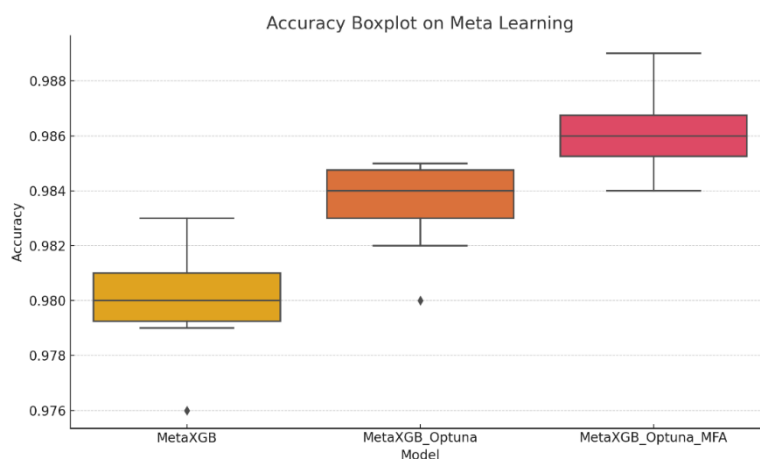


Figure 8 Accuracy Distribution of Meta Learning Models using Boxplot Visualization

Figure 8 shows that the MetaXGB_Optuna_MFA model produces the highest and most consistent accuracy distribution compared to the other two variants. The median accuracy of this model is around 0.986, with a narrow interquartile range and no significant outliers, indicating high predictive stability. Meanwhile, the MetaXGB model has the lowest median (~0.980) with a wider spread of values and the presence of outliers, indicating greater fluctuations in performance. The addition of Optuna to the MetaXGB model was shown to improve median accuracy (~0.984) and reduce variation, but the most significant improvement occurred when MFA was applied simultaneously. These findings confirm that the combination of hyperparameter tuning with Optuna and MFA can synergistically strengthen model generalization and reliability in meta learning scenarios.

To evaluate the generalization ability of the model in more detail, the confusion matrix and ROC curves of three meta models at Layer 3 were analyzed: MetaXGB, MetaXGB optimized with Optuna, and MetaXGB combined with Optuna and MFA. The results of the multiclass classification evaluation of the four risk classes are shown in Figure 9.

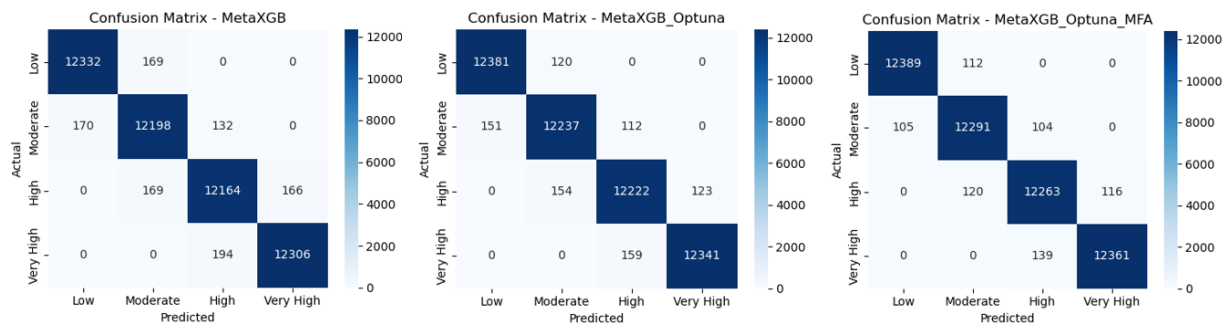


Figure 9 Confusion Matrix Comparison of Meta Models

The confusion matrix analysis in Figure 9 shows that the integration of Optuna and MFA progressively reduces the number of misclassifications between classes. The MetaXGB model only correctly classified the “Low” class for 12,332 samples, while the MetaXGB_Optuna_MFA model increased to 12,389. The same is seen for the other classes, where the number of correct predictions increases with the MFA approach. Figure 10 corroborates these findings through the ROC curves, which show AUC = 1.00 for all classes across all three models. Although all models recorded perfect AUCs, the improved performance is still reflected in the more precise distribution of predictions in the confusion matrix. This indicates that the combination of Optuna and MFA not only improved the average accuracy, but also refined the model's sensitivity to all classes proportionally.

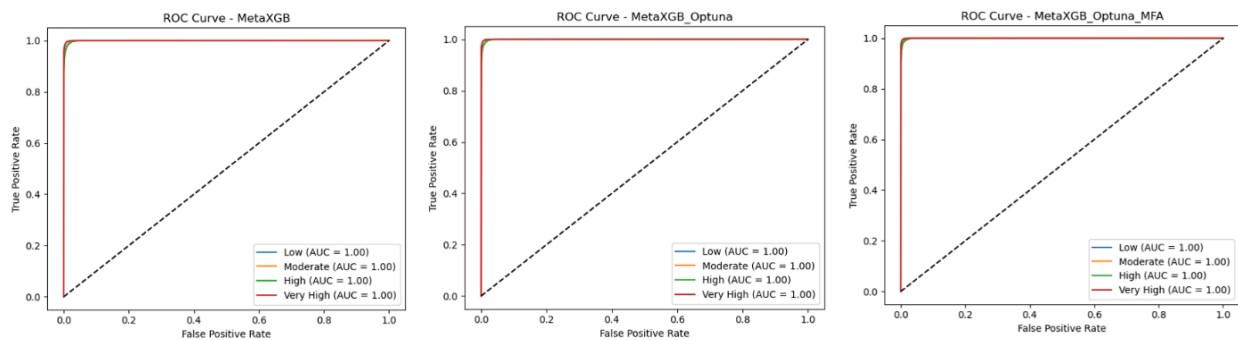


Figure 10 ROC Curve Evaluation of Meta Models

To ensure that the model does not experience overfitting and has consistent generalization ability to new data, a 10-Fold Cross Validation scheme was used. This evaluation tests the stability of the model in various subsets of training and test data, resulting in performance metrics at each fold as presented in the following Table 4.

Table 4 Cross Validation Performance of XGB + Optuna + MFA Model

Fold	Accuracy (%)	Precision	Recall	F1-Score
Fold 1	98.40	0.98	0.98	0.98
Fold 2	98.55	0.99	0.99	0.98
Fold 3	98.67	0.99	0.99	0.99
Fold 4	98.79	0.99	0.99	0.99
Fold 5	98.91	0.99	0.99	0.99
Fold 6	98.62	0.98	0.99	0.98
Fold 7	98.73	0.99	0.98	0.99
Fold 8	98.85	0.99	0.99	0.99
Fold 9	98.70	0.99	0.98	0.98
Fold 10	98.78	0.99	0.99	0.99
Average	98.69	0.99	0.99	0.99

The cross-validation results show that the XGB + Optuna + MFA model has a very stable performance with minimal variation in metrics between folds. Accuracy ranged from 98.40% to 98.91%, with an average of 98.69%. Precision, recall, and F1-score also showed high consistency with an overall average of 0.99. This indicates that the integration of hyperparameter optimization (Optuna) and meta feature augmentation (MFA) techniques not only improved model performance on the training data, but also produced a model that is generalizable and robust to data variations. There is no indication of overfitting as the performance between folds is relatively balanced and does not show extreme deviations. The Recall values for the High and Very High classes are consistently high, indicating that the MLSEL model is effective at identifying high risk flood conditions. This demonstrates that the performance is not driven by overall accuracy alone but also by strong sensitivity to high risk events, as further supported by the Confusion Matrix in Figure 10.

4. Discussion

Visualizing the accuracy comparison between MLSEL models helps highlight the contribution of each layer as well as the effect of integrating feature optimization and augmentation techniques. The accuracy comparison of all models can be seen in the following Figure 11.

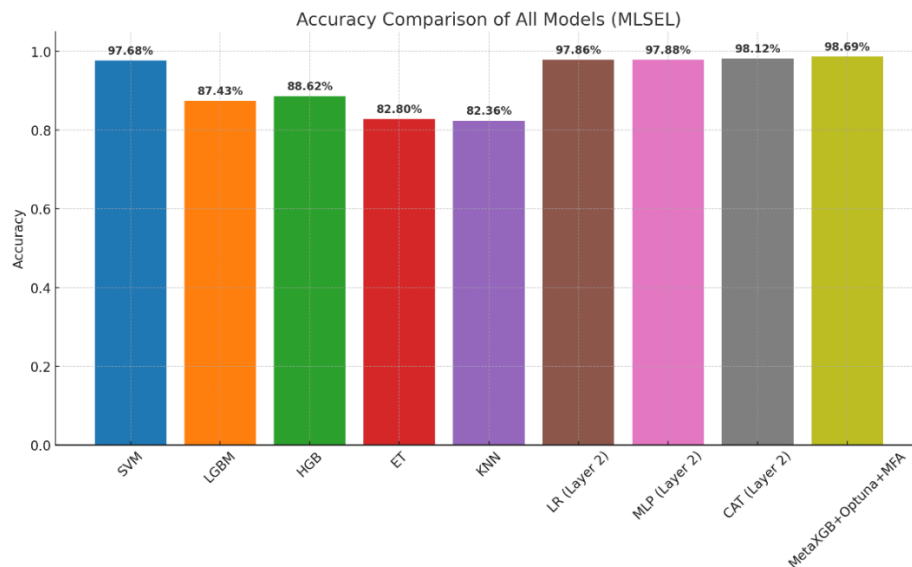


Figure 11 Accuracy Comparison Across Base Learners and Meta-Stacked Models Using MLSEL Architecture

The visualization results show a consistent increase in accuracy as the complexity of the MLSEL architecture increases. Baseline models such as KNN and ET have the lowest accuracy, at 82.36% and 82.80% respectively. While SVM as the best single ML model recorded a high accuracy of 97.68%, it was still lower than all models in Layer 2 (MLP, LR, and CatBoost) which ranged from 97.86% to 98.12%. The highest performing model was XGB + Optuna + MFA at Layer 3, with an accuracy of 98.69%. This improvement shows that the combination of boosting techniques, hyperparameter tuning (Optuna), and Meta Feature Augmentation can enrich information representation and minimize misclassification significantly, making this model superior in the MLSEL framework.

To ensure the superiority of the proposed model, a performance comparison was conducted against several stacked ensemble learning (SEL) models from previous studies that used the same dataset as this study. This evaluation reflects not only accuracy, but also accuracy and balance in classification between classes. The performance comparison between the MLSEL model and previous approaches is shown in Table 5.

Table 5 Comparative Evaluation of MLSEL and Previous Researchers SEL Models on the Same Dataset

Measures	Accuracy	Precision	Recall	F1-Score
SEL 1 [35]	94.82	0.95	0.94	0.94
SEL 2 [31]	96.21	0.96	0.96	0.96
SEL 3 [36]	95.37	0.95	0.95	0.95
SEL 4 [37]	97.31	0.97	0.97	0.97
SEL 5 [38]	96.89	0.96	0.96	0.96
MLSEL	98.69	0.99	0.99	0.99

Table 5 shows that the MLSEL model consistently outperforms the previous five SEL approaches on all evaluation metrics. The best model from the previous study (SEL 4) only achieved 97.31% accuracy, precision 0.97, recall 0.97, and F1-score 0.97. In contrast, MLSEL produced the highest scores on all four key metrics, namely 98.69% accuracy, 0.99 precision, 0.99 recall, and 0.99 F1-score. This performance spike indicates that the multi-layer architecture, hyperparameter optimization (Optuna), and MFA synergistically improve the generalization capability of the model.

In order to evaluate the robustness and generalizability of the proposed model, a comparison is made to several previous studies using different datasets, namely disaster classification. Although the data context is different, this comparison is important to test whether the multi-layer structure and optimization used in MLSEL are able to maintain high performance and avoid overfitting. Details of MLSEL performance compared to previous models are presented in Table 6.

Table 6 Comparative Performance of MLSEL and Existing Models on Disaster Classification Dataset

Measures	Accuracy	Precision	Recall	F1-Score
SEL 1 [39]	94.21	0.93	0.93	0.93
SEL 2 [40]	95.05	0.94	0.94	0.94
SEL 3 [41]	96.37	0.95	0.95	0.95
SEL 4 [42]	97.01	0.96	0.96	0.96
SEL 5 [43]	96.88	0.96	0.96	0.96
MLSEL	97.85	0.97	0.97	0.97

Table 6 shows that MLSEL still maintains the highest performance among all previous models despite being applied to different domains. The best model from the previous study (SEL 4) only achieved 97.01% accuracy with precision, recall, and F1-score of 0.96. SEL 4 corresponds to a stacking ensemble model that integrates Random Forest, Gradient Boosting, Support Vector Machine, and K Nearest Neighbors as base learners with Logistic Regression as a meta-classifier. However, SEL 4 relies on a single-layer stacking structure without meta-feature augmentation or automatic hyperparameter optimization, which limits its representation capacity compared to the multi-layer MLSEL framework proposed in this study. In comparison, MLSEL was able to achieve 97.85% accuracy and a perfect score of 0.97 for precision, recall, and F1-score. This indicates that the proposed multilayer stacking architecture not only works optimally on the original domain of the dataset, but can also be effectively transferred to other classification problems with competitive results. The dominance of MLSEL's quantitative scores confirms the stability and effectiveness of the integrative approach through hyperparameter optimization using Optuna and MFA.

To evaluate the computational efficiency, we compared the Training Time and Inference Time of the MLSEL framework with the tuned baseline models. The MLSEL model required a longer training duration because it includes three sequential learning layers and an Optuna based optimization process. However, the inference time remained efficient and only slightly higher than the baseline models since prediction is executed through a forward pass across the stacked learners. The observed increase in training cost is justified by the proportional improvement in predictive performance, including higher accuracy, stronger recall for high risk classes, and reduced variance across folds, demonstrating that the additional model depth provides meaningful benefits beyond the simpler baseline configurations.

Recent studies have also highlighted the growing role of ensemble learning in modeling climate related hazards. For example ensemble based frameworks have been used to enhance climate projections using CMIP6 data for the Middle East and to assess future drought risks in central and southern regions of the Middle East [44]. These works show that ensemble architectures can effectively capture complex environmental variability and improve long term risk forecasting. The findings from the proposed MLSEL model are consistent with these broader advances, demonstrating that multilayer ensemble learning provides strong predictive capability for flood risk classification in hydrologically dynamic regions.

5. Conclusions

This study proposed a MLSEL model that integrates base-level learners, optimized meta learners, and MFA to enhance multiclass classification performance. Experimental results on datasets demonstrate that the proposed MLSEL framework consistently outperformed conventional single-layer and baseline ensemble models across all evaluation metrics. The best-performing configuration, XGB + Optuna + MFA, achieved an average accuracy of 98.69%, along with 0.99 precision, recall, and F1-score, indicating exceptional robustness and generalization. Further comparative analyses against state of the art methods on the same and different datasets reaffirmed the superior predictive capability of MLSEL. Its consistent dominance across multiple folds and tasks highlights the strength of deep meta-learning integration, especially when combined with advanced hyperparameter tuning and feature augmentation. For future research, it is recommended to incorporate weighted base learners, where accuracy-based or confidence-based weighting schemes could further refine the aggregation process at the meta level, potentially leading to even higher performance and interpretability across diverse classification domains.

6. References

- [1] Zhai S, Bartkowiak N, Sibirtsev S, Jupke A. Experimental determination and model-based prediction of flooding points in a pilot-scale continuous liquid-liquid gravity separator. *Sep Purif Technol.* 2025;377:134177.
- [2] Bian W, Fang J, Wang P, Sun Q, Fang J. Deep learning surrogate models for spatiotemporal prediction of coastal flooding inundations in Tianjin, China. *J Hydrol: Reg Stud.* 2025;60:102593.
- [3] Alam MG, Tripathi V, Bhatt CM, Mohanty MP. A novel framework embedding Bayesian-optimized ensemble machine learning and explainable artificial intelligence (XAI) to improve flood prediction in complex watersheds. *Environ Sustain Indic.* 2025;27:100760.
- [4] Jang SD, Yoo JH, Lee YS, Kim B. Flood prediction in urban areas based on machine learning considering the statistical characteristics of rainfall. *Prog Disaster Sci.* 2025;26:100415.
- [5] Antwi-Agyakwa KT, Afenyo MK, Angnuureng DB. Know to predict, forecast to warn: a review of flood risk prediction tools. *Water.* 2023;15(3):427.
- [6] Tian J, Yan Y, Zeng S. Intelligent identification and management of flood risk areas in high-density blocks from the perspective of flood regulation supply and demand matching. *Ecol Indic.* 2024;160:111799.
- [7] Wang X, Liu R, Sun C, Zhai X, Ding L, Liu X, et al. Optimizing flood resilience in China's mountainous areas: design flood estimation using advanced machine learning techniques. *J Hydrol: Reg Stud.* 2025;59:102345.
- [8] Tanhapour M, Soltani J, Shakibian H, Malekmohammadi B, Hlavcova K, Kohnova S, et al. The enhanced integration of proven techniques to quantify the uncertainty of forecasting extreme flood events based on numerical weather prediction models. *Weather Clim Extrem.* 2025;48:100767.
- [9] Joubier V, Ebtehaj I, Amiri A, Gumiere SJ, Bonakdari H. Multitemporal river flow discharge prediction: a new framework for integrated environmental management and flood control. *J Environ Manage.* 2025;383:125372.
- [10] Islam MT, Roy SC, Jahan N, Al-Mahmud, Islam MM, Ferdous AA, et al. Comparative evaluation of machine learning models for extreme river water level forecasting in Bangladesh: Implications for flood and drought resilience. *Prog Disaster Sci.* 2025;27:100449.
- [11] Nazli S, Liu J, Song T, Soomro S, Wang H. A hybrid machine learning approach to unravel monsoon variability and meteorological dynamics of Pakistan's 2010 and 2022 historic floods. *J Hydrol: Reg Stud.* 2025;60:102505.
- [12] Asfaw W, Rientjes T, Bekele TW, Haile AT. Estimating elements susceptible to urban flooding using multisource data and machine learning. *Int J Disaster Risk Reduct.* 2025;116:105169.
- [13] Lubis A, Irawan Y, Junadhi J, Defit S. Leveraging K-Nearest neighbors with SMOTE and Boosting techniques for data imbalance and accuracy improvement. *J Appl Data Sci.* 2024;5(4):1625-38.

- [14] Yang PB, Chan YJ, Yazdi SK, Lim JW. Optimisation and economic analysis of industrial-scale anaerobic co-digestion (ACoD) of palm oil mill effluent (POME) and decanter cake (DC) using machine learning models: a comparative study of Gradient Boosting Machines (GBM), K-nearest neighbours (KNN). *J Water Process Eng.* 2024;58:104752.
- [15] Yousefi M, Oskoei V, Esmaili HR, Baziar M. An innovative combination of extra trees within adaboost for accurate prediction of agricultural water quality indices. *Results Eng.* 2024;24:103534.
- [16] Wei X, Xu Y, Li X, Fan G, Cheng X, Yu T, et al. Study on prediction model of nitrogen oxide concentration in reprocessing plant based on random forest. *Int J Adv Nucl React Des Technol.* 2025;7:63-9.
- [17] Farooq O, Shahid M, Arshad S, Altaf A, Iqbal F, Vera YAM, et al. An enhanced approach for predicting air pollution using quantum support vector machine. *Sci Rep.* 2024;14:19521.
- [18] Zheng H, Sherazi SWA, Lee JY. A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data. *IEEE Access.* 2021;9:113692-704.
- [19] Lin S, Nong X, Luo J, Wang C. A novel multi-model stacking ensemble learning method for metro traction energy prediction. *IEEE Access.* 2022;10:129231-44.
- [20] Song D, Yi T, Xiang Q, Chen H. Leveraging ISMOTE-KPCA-STACKING Algorithm for enhanced vascular vertigo / dizziness diagnosis and clinical decision support. *IEEE Access.* 2023;11:99734-51.
- [21] Garouani M, Barhrhouj A, Teste O. XStacking : an effective and inherently explainable framework for stacked ensemble learning. *Inf Fusion.* 2025;124:103358.
- [22] Mostofi S, Yilmaz Z, Başağa HB, Okur FY, Altunişik AC, Taciroglu E. A hybrid stacked ensemble model for rapid seismic damage assessment with imbalanced training data: a case study on the 2023 Kahramanmaraş earthquakes. *Eng Struct.* 2025;340:120754.
- [23] Chen X, Liu J, Wu C. Multi-class financial distress prediction based on hybrid feature selection and improved stacking ensemble model. *Expert Syst Appl.* 2025;282:127832.
- [24] Shafieian S, Zulkernine M. Multi-layer stacking ensemble learners for low footprint network intrusion detection. *Complex Intell Syst.* 2023;9:3787-99.
- [25] Li M, Chen L. An interpretable and stacking ensemble model for predicting heat and mass transfer of desiccant wheel. *Results Eng.* 2025;25:104181.
- [26] Naveen Venkatesh S, Sripada D, Sugumarna V, Aghaei M. Detection of visual faults in photovoltaic modules using a stacking ensemble approach. *Heliyon.* 2024;10(6):e27894.
- [27] Shishodia V, Singh V, Thampi SG. Implementation of stack-based ensemble technique for classification of glaciers in the western Himalayan catchments. *Phys Chem Earth, Parts A/B/C.* 2024;136:103723.
- [28] Li Y, Li K, Wang K, Li G, Wang Z. An improved stacking model for forest fire susceptibility prediction in Chongqing City, China. 2023 IEEE Smart World Congress (SWC); 2023 Aug 28-31; Portsmouth, United Kingdom. USA: IEEE; 2023. p. 1-6.
- [29] Hoc HT, Silhavy R, Prokopova Z, Silhavy P. Comparing stacking ensemble and deep learning for software project effort estimation. *IEEE Access.* 2023;11:60590-604.
- [30] Mota LFM, Giannuzzi D, Bisutti V, Pegolo S, Trevisi E, Schiavon S, et al. Real-time milk analysis integrated with stacking ensemble learning as a tool for the daily prediction of cheese-making traits in Holstein cattle. *J Dairy Sci.* 2022;105(5):4237-55.
- [31] Herianto, Kurniawan B, Hartomi ZH, Irawan Y, Anam MK. Machine learning algorithm optimization using stacking technique for graduation prediction. *J Appl Data Sci.* 2024;5(3):1272-85.
- [32] Cawood P, van Zyl TL. Feature-weighted stacking for nonseasonal time series forecasts: a case study of the COVID-19 epidemic curves. 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI); 2021 Nov 26-27; Cairo, Egypt. USA: IEEE; 2021. p. 53-9.
- [33] Devis Y, Muhamadiyah, Yulanda, Irawan Y, Wahyuni R. Optimization of machine learning models for risk prediction of DHF spread to support management strategies in urban areas. *J Appl Data Sci.* 2025;6(4):2407-20.
- [34] Muhaimin A, Edriyansyah, Wahyat, Irawan Y, Wahyuni R. Optimized IoT-based multimodal fusion for early forest fire detection and prediction. *ECTI Trans Comput Inf Technol.* 2025;19(4):569-82.
- [35] Fonda H, Irawan Y, Melyanti R, Wahyuni R, Muhaimin A. A comprehensive stacking ensemble approach for stress level classification in higher education. *J Appl Data Sci.* 2024;5(4):1701-14.
- [36] Alalwany E, Alsharif B, Alotaibi Y, Alfahaid A, Mahgoub I, Ilyas M. Stacking ensemble deep learning for real-time intrusion detection in IoMT environments. *Sensors.* 2025;25(3):624.
- [37] Gu J, Liu S, Zhou Z, Chalov SR, Zhuang Q. A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. *Water.* 2022;14(3):492.
- [38] Hassanali M, Soltanaghaei M, Javdani Gandomani T, Zamani Boroujeni F. Exploring stacking methods for software effort estimation with hyperparameter tuning. *Cluster Comput.* 2025;28:241.
- [39] Alqahtani AF, Ilyas M. An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. *Mach Learn Knowl Extr.* 2024;6(1):156-70.
- [40] Megouo TGP, Pierre S. A stacking ensemble machine learning model for emergency call forecasting. *IEEE Access.* 2024;12:115820-37.
- [41] Öz E, Bulut O, Cellat ZF, Yürekli H. Stacking: an ensemble learning approach to predict student performance in PISA 2022. *Educ Inf Technol.* 2025;30:7753-79.
- [42] Kalabarige LR, Rao RS, Pais AR, Gabralla LA. A boosting-based hybrid feature selection and multi-layer stacked ensemble learning model to detect phishing websites. *IEEE Access.* 2023;11:71180-93.
- [43] Abbas S, Sampedro GA, Abisado M, Almadhor AS, Kim TH, Zaidi MM. A novel drug-drug indicator dataset and ensemble stacking model for detection and classification of drug-drug interaction indicators. *IEEE Access.* 2023;11:101525-36.
- [44] Khosravi Y, Ouarda TBMJ, Homayouni S. Developing an ensemble machine learning framework for enhanced climate projections using CMIP6 data in the Middle East. *Npj Clim Atmos Sci.* 2025;8:174.