JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 8, No. 2, December 2025, pp. 650~661
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v8i2.4639

An Integrated Machine Learning and Deep Learning Approach for Multiclass Flood Risk Classification with Feature Selection and Imbalanced Data Handling

Yuda Irawan[1], Refni Wahyuni[2], Herianto[3], Muhammad Habib Yuhandri[4]
[1,2]*Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*
[3]Information System, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia
[4]Information System, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

Article Info

Article history:

Received mm dd, yyyy
Revised mm dd, yyyy
Accepted mm dd, yyyy

Keywords:

Machine Learning
Deep Learning
Flood Risk Prediction
SMOTEENN
LASSO

ABSTRACT

Floods are hydrometeorological disasters that often occur in tropical regions such as Indonesia and can have significant impacts on infrastructure, economy, and public health. This study aims to build and compare the performance of 21 artificial intelligence models, consisting of 15 Machine Learning algorithms and 6 Deep Learning architectures, in classifying flood risk levels based on multivariate tabular data. The dataset used includes 22 relevant environmental and social variables, with classification targets in four classes: Low, Moderate, High, and Very High. To improve data quality, feature selection was carried out using the LASSO method and class balancing with the SMOTEENN technique. The evaluation results showed that the C4.5, MLP, Random Forest, and Logistic Regression models obtained the highest accuracy (>94%), followed by deep learning models such as BiLSTM, CNN, and BiGRU with competitive accuracy (≥90%). Confusion matrix analysis confirmed the consistency of predictions across classes with a balanced distribution, especially in the decision tree and deep neural network models. This study emphasizes the importance of selecting a model that suits the characteristics of the data to achieve optimal predictions. The pipeline developed in this study is expected to be the basis for a more accurate and adaptive AI-based early warning system in mitigating flood risks in the future.

Corresponding Author:

Yuda Irawan
Department of Computer Science
Universitas Hang Tuah Pekanbaru
Pekanbaru, Indonesia
Email: yudairawan89@mail.com

1. Introduction

Floods are one of the most frequent hydrometeorological disasters in Indonesia and other tropical countries[1]. The impacts not only include damage to infrastructure and the economy, but also loss of life and disruption to ecosystems and public health[2]. The complexity of flood causes that include natural and anthropogenic factors demands an intelligent data-based approach to predict the level of risk accurately and adaptively[3].

With the increasing availability of structured environmental and social data, artificial intelligence (AI) approaches through machine learning (ML) and deep learning (DL) have great potential in building flood risk prediction systems[4,5]. In various studies, ML models such as Decision Tree, Random Forest, LightGBM, Gradient Boosting, and XGBoost have been shown to provide high accuracy on tabular data, mainly due to their ability to capture non-linear relationships and their ease of interpretation[6–10].

Previous studies have widely utilized ML for flood prediction based on tabular data. Models such as Decision Tree, Random Forest, LightGBM, Gradient Boosting, and XGBoost have become popular approaches due to their stability, ability to handle multivariate numerical data, and relatively simple model interpretation[11–15].

Deep learning models such as LSTM, BiLSTM, CNN, and Transformer, although commonly used for time-series and spatial data, are also starting to be explored for structured data-based classification[16–19]. Deep learning approaches such as LSTM, BiLSTM, CNN, and Transformer are widely used for sequential or spatial data, such as satellite imagery and text data[20–22]. The results of studies by several researchers show that DL still faces challenges in outperforming ML on ordinary tabular data, unless the data is very large or has complex sequential patterns[23,24]. Nevertheless, the DL model remains relevant to be studied as an alternative classification of flood risk levels if given an appropriate architecture[25].

Several studies have shown that DL models do not necessarily outperform ML models when applied to tabular data, especially in the context of predictive classification such as flood risk. Others have concluded that DL models tend to overfit and require complex tuning, and are often outperformed by simple, well-tuned tree-based models for structured data[26]. In addition, most studies only focus on one type of model or use a suboptimal pipeline, making the evaluation results difficult to compare fairly.

The dataset used in this study consists of 22 features covering indicators such as Monsoon Intensity, Urbanization, Deforestation, Siltation, Drainage Systems, to Political Factors. The classification target is Flood Risk Level which is divided into four classes: Low, Moderate, High, and Very High. With these data characteristics, a multi-class classification approach based on structured data is a challenge as well as an opportunity to comprehensively evaluate the performance of various AI models.

However, two main issues that often arise in disaster classification data are class imbalance and feature redundancy[27]. To overcome this, this study applies two important methods: SMOTEENN and LASSO Regression. SMOTEENN is a combination of Synthetic Minority Oversampling Technique and Edited Nearest Neighbors, which not only balances the amount of data between classes, but also cleans noisy data from the majority class[28]. This method is proven to be superior to regular SMOTE in improving model generalization in multi-class classification cases[29].

LASSO (Least Absolute Shrinkage and Selection Operator) is used as an effective feature selection technique to filter important features by penalizing small regression coefficients[30,31]. The use of LASSO helps prevent overfitting and increases model efficiency, especially on data with many numerical features such as in this study[32].

The main novelty in this study is the integration of SMOTEENN and LASSO-based preprocessing pipelines, which are then used to systematically compare the performance of 15 machine learning models and 6 deep learning models in classifying flood risk levels based on multivariate structured data. This dual-stage preprocessing approach combines resampling and feature selection, enabling the dataset to become both more balanced and more informative before being fed into the classification models. By applying SMOTEENN, the study addresses class imbalance through synthetic oversampling while simultaneously removing noisy instances, resulting in cleaner and more representative training data. LASSO further refines the dataset by selecting the most relevant predictors, reducing dimensionality and preventing overfitting.

Beyond providing performance evaluation results using common metrics such as accuracy, precision, recall, and F1-score, this research also examines how each preprocessing stage contributes to improvements in model robustness and prediction consistency. The comparative analysis offers deeper insights into which algorithms benefit most from the enhanced data pipeline and how the combination of SMOTEENN and LASSO influences different model families. Overall, this study contributes a comprehensive framework for improving flood risk classification, demonstrating that thoughtful preprocessing can significantly elevate predictive performance across diverse modeling approaches.

2. Research Method

This research was conducted through a series of systematic and structured stages to obtain an optimal flood risk prediction model based on structured data. Each stage is designed to ensure that the data used has gone through the right selection, balancing, and modeling process, so that the final results obtained can be relied upon for the purpose of multi-class flood risk classification. The process starts from data collection,

followed by preprocessing, feature selection, to training and evaluation of machine learning and deep learning models. The research flow can be explained through the following figure:

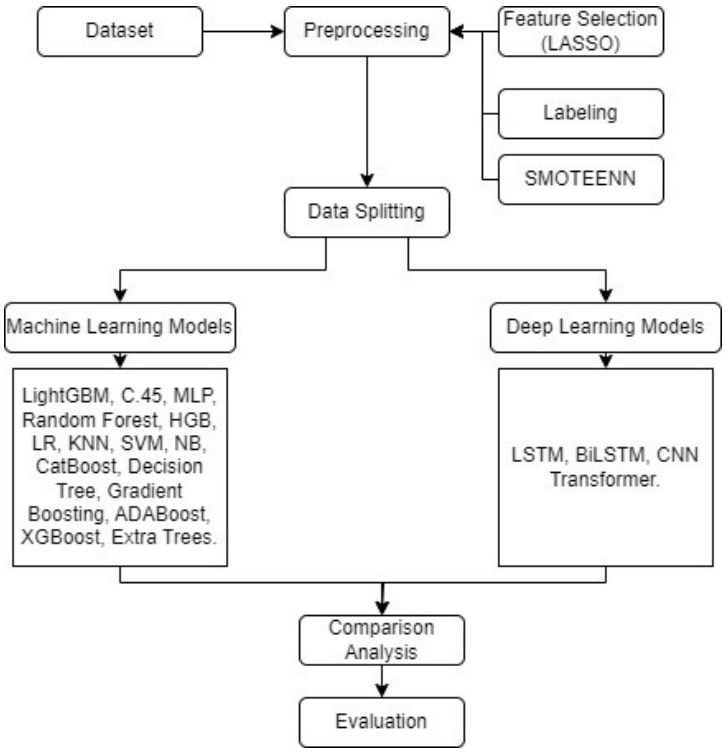

Figure 1. Research Flow Diagram

The stages of the research can be explained as follows:

1. Dataset Collection

The initial stage of this research is the process of collecting data used to build a flood risk prediction model. The data that has been collected amounts to 50,000 data, which includes various environmental and social features such as monsoon rainfall intensity, river management quality, deforestation, urbanization, drainage systems, and political factors. This dataset is structured in tabular format and has been successfully collected as the main foundation for the next research process. The dataset was obtained from Kaggle.com, an open-access data repository, where the information is publicly available and licensed for research use. All data used in this study are anonymized and aggregated, containing no personal identifiers. Therefore, no additional ethical clearance was required, and the use of the dataset complies with the terms and conditions of Kaggle's open data license.

2. Preprocessing

After the data has been successfully collected, the next stage that will be carried out is preprocessing. This stage includes checking for duplicate data and missing values, as well as standardization or normalization of data if necessary. In addition, adjustments to the variable format and encoding of categorical features (if any) will be carried out so that the data is compatible with the machine learning and deep learning algorithms used.

3. Feature Selection using LASSO

To simplify the model and improve prediction accuracy, this study will apply the Least Absolute Shrinkage and Selection Operator (LASSO) method. This method functions to select the most relevant and significant features to the target variable, while eliminating redundant or less influential features. Thus, the model built becomes more efficient and avoids overfitting.

4. Labeling

The labeling stage will be carried out to determine the flood risk class for each data based on certain parameters available in the dataset. The labels used are multi-class, consisting of four categories: Low, Moderate, High, and Very High. This labeling process is important for defining the target variable in the supervised learning process in classification.

5. Data Balancing

Because the class distribution in the dataset is likely to be unbalanced, the SMOTEENN technique will be applied. This method is a combination of SMOTE (Synthetic Minority Oversampling Technique) to add synthetic data to the minority class, and ENN (Edited Nearest Neighbor) to remove noisy data from the majority class. The goal is that the model built is not biased towards the majority class and is able to recognize patterns from the minority class better.

6. Data Splitting

After the data has been preprocessed and balanced, the next step is to separate the data into two parts: training data and testing data. This division will be done with a general ratio such as 80:20 or 70:30. Training data will be used to build models, while testing data will be used to evaluate model performance objectively and independently.

7. Machine Learning and Deep Learning Model Training

At this stage, training data will be used to build and train prediction models using various algorithms. For machine learning, 15 machine learning models will be used. Meanwhile, for deep learning, the LSTM, BiLSTM, CNN, RNN, GRU, and BiGRU architectures will be used. These models will be compared based on the results of the flood risk classification.

8. Comparative Analysis

After all models have been trained and tested, a comparative analysis will be carried out to find out which model gives the best results. This analysis will be carried out by comparing the performance of each model based on evaluation metrics such as accuracy, precision, recall, and F1-score, as well as the Confusion Matrix. With this analysis, the advantages and disadvantages of each approach can be identified quantitatively.

9. Evaluation

The final stage of this research is a comprehensive evaluation of the results of the model and the process that has been carried out. The evaluation includes not only classification performance, but also the effectiveness of preprocessing (SMOTEENN and LASSO), computational efficiency, and potential implementation in the field. From the results of this evaluation, recommendations will be given for the best prediction model that is suitable for application to structured flood data in real scenarios.

3. Result and Discussion

3.1. Dataset and Preprocessing

The dataset used in this study consists of 50,000 entries with 22 predictor features and 1 classification target, all organized in a structured tabular format. Each predictor feature represents a quantitative indicator related to flood risk, covering environmental, infrastructural, and socio-economic dimensions. Key variables include river governance quality, levels of deforestation, rates of urbanization, drainage system performance, rainfall intensity, soil permeability, land-use patterns, and population density. These features were selected to ensure a comprehensive representation of the factors that influence flood vulnerability across different regions. The dataset's large size supports robust model training and reliable

evaluation, enabling both traditional machine learning and deep learning models to capture meaningful patterns. An illustration of the dataset structure and feature distribution is presented in Figure 2, providing a clearer overview of the data components used throughout the analysis.

TopographyDrainage	RiverManagement	Deforestation	Urbanization	ClimateChange	DamsQuality	Siltation	AgriculturalPractices	Encroachments	IneffectiveDisasterPreparedness	DrainageSystems	CoastalVulnerability	Landslides	Watersheds	DeterioratingInfrastructure	PopulationScore	WetlandLoss	InadequatePlanning	PoliticalFactors	Flood Risk Level	Flood Probability	id Flood Risk Level
8	6	6	4	4	6	2	3	2	5	10	7	4	2	3	4	3	2	6	High	0,312	2
4	5	7	7	9	1	5	5	4	6	9	2	6	2	1	1	9	1	3	Very High	0,357	3
10	4	1	7	5	4	7	4	9	2	7	4	4	8	6	1	8	3	6	Very High	0,324	3
4	2	7	3	4	1	4	6	4	9	4	2	6	6	8	8	6	6	10	Moderate	0,271	1
7	5	2	5	8	5	2	7	5	7	7	6	5	3	3	4	4	3	4	Moderate	0,259	1
6	6	4	6	4	3	1	3	5	1	10	5	9	5	5	7	3	3	2	Very High	0,336	3
7	4	5	5	5	4	8	8	4	6	8	4	5	4	7	7	5	4	8	Very High	0,339	3
3	5	5	6	6	6	7	6	5	5	4	6	9	7	10	6	5	4	5	Very High	0,327	3
3	5	4	5	11	3	2	9	7	8	2	8	7	5	4	9	6	5	7	Low	0,251	0
3	5	6	2	3	7	7	10	4	5	7	6	5	6	7	5	7	4	8	High	0,293	2
1	7	4	5	7	4	3	0	2	6	6	4	8	5	5	7	4	2	6	Moderate	0,27	1
9	1	4	3	7	5	8	4	6	5	8	4	3	5	6	4	6	14	3	High	0,306	2
9	4	1	5	4	2	8	4	5	4	5	7	7	3	4	2	3	6	3	Low	0,252	0
3	7	9	7	4	11	7	8	3	1	4	5	4	2	3	4	6	7	4	Very High	0,326	3
1	9	4	6	7	6	3	4	2	8	4	3	2	6	4	5	4	2	8	High	0,295	2

Figure 2. Initial Dataset View of Flood Risk Prediction

Based on Figure 2, each feature is represented in the form of a standardized numeric value on an ordinal scale range between 1 and 10. The Flood Risk Level feature shows a qualitative class label (Low, Moderate, High, and Very High), while the Flood Risk Level id column is used as a numeric target for the purposes of machine learning and deep learning model classification. In addition, the Flood Probability feature is a continuous numeric and is not used as the main target. The distribution pattern shows that not all classes have a balanced amount of data, so a balancing technique is needed to avoid classification bias towards the majority class.

3.2. Feature Selection

The feature selection process was carried out using the LASSO (Least Absolute Shrinkage and Selection Operator) method, which effectively filters the most significant features for the flood risk classification target. LASSO penalizes small regression coefficients, so that less relevant features are automatically eliminated. The results of this process produce eight main features that are retained, namely: MonsoonIntensity, TopographyDrainage, RiverManagement, Deforestation, Urbanization, DrainageSystems, WetlandLoss, and DeterioratingInfrastructure.

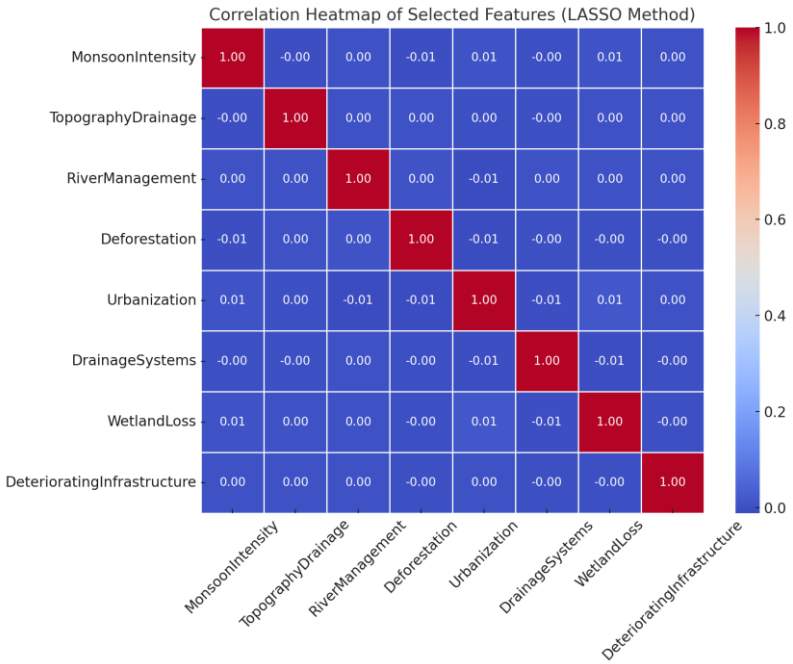

Figure 3. Heatmap of Correlation Between Selected Features (LASSO)

Figure 3 shows the correlation heatmap between features selected using the LASSO method. The results of the correlation analysis indicate that most features have very low correlation with each other, with

coefficient values ranging from -0.01 to 0.01, except for the main diagonal which represents a perfect correlation (value 1.00) with itself. This indicates that there is no significant multicollinearity between features, so that each feature contributes unique information to the classification model. The diversity of information between features is important in building a prediction model that is not only accurate but also resistant to overfitting, especially in complex tabular data such as in the case of flood risk prediction.

3.3. Class Distribution and Data Balancing

The class distribution in the original dataset shows an imbalance in the number of samples between flood risk classes. This imbalance has the potential to cause the classification model to be biased towards the majority class and fail to recognize patterns from the minority class effectively. To overcome this, two data balancing approaches were used: SMOTE and SMOTEENN. The class distribution display before and after balancing can be seen in Figure 4.

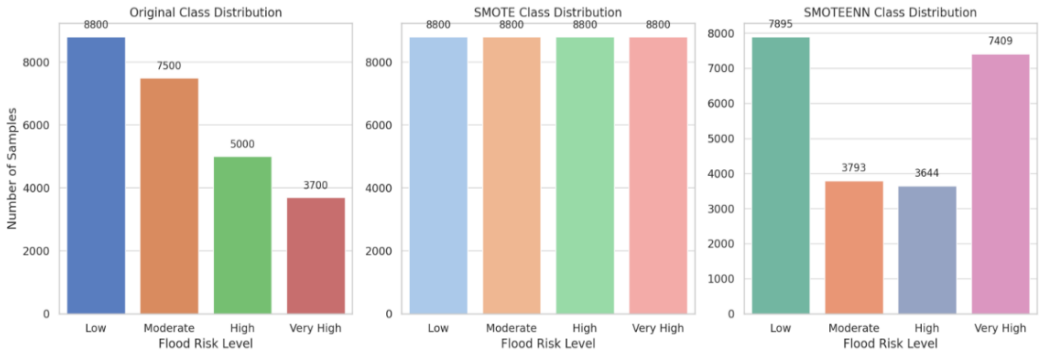

Figure 4. Comparison of Class Distribution: Original, SMOTE, and SMOTEENN

Based on Figure 4, it can be seen that the SMOTE method produces a truly balanced class distribution, with the same number of samples for all four classes. However, this approach tends to add synthetic data without considering the possibility of noisy or borderline data, which can increase the risk of overfitting, especially in minority classes that were previously very limited. In contrast, the SMOTEENN method provides more adaptive results. Although the class distribution is not completely balanced, this approach combines the advantages of oversampling from SMOTE and the ability of ENN to remove noisy or ambiguous data from the majority class. The end result is a distribution that is more representative of real patterns, with reduced potential noise and increased quality of model generalization. Therefore, SMOTEENN is considered more effective in maintaining the balance and cleanliness of training data for flood risk classification.

3.4. Machine Learning Model Evaluation Results

To evaluate the performance of each machine learning model in classifying flood risk levels, 15 algorithms were tested using the main evaluation metrics of accuracy, precision, recall, and F1-score. Each model was trained on LASSO and SMOTEENN preprocessing data to ensure efficiency and balance of class distribution. The classification results can be seen in Table 1.

Table 1. Classification Report of Machine Learning Models on Flood Risk Prediction Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.94	0.95	0.94	0.94
MLP	0.96	0.96	0.97	0.96
LightGBM	0.94	0.95	0.95	0.95
Random Forest	0.95	0.95	0.95	0.95
C4.5	0.97	0.96	0.96	0.96
HistGradientBoosting	0.93	0.93	0.92	0.92
SVM	0.77	0.77	0.77	0.77
XGBoost	0.76	0.76	0.76	0.76
Extra Trees	0.72	0.71	0.72	0.69
KNN	0.71	0.69	0.71	0.69
AdaBoost	0.7	0.8	0.7	0.7

Gradient Boosting	0.7	0.7	0.7	0.68
Naive Bayes	0.68	0.67	0.68	0.64
CatBoost	0.61	0.6	0.61	0.6
Decision Tree	0.6	0.59	0.6	0.6

The results in Table 1 show that the top six models — namely Logistic Regression, MLP, LightGBM, Random Forest, C4.5, and HistGradientBoosting — provide very high classification performance, with accuracies ranging from 93% to 97%. The C4.5 and MLP models even record accuracy and F1-score values of up to 96%–97%, reflecting their ability to capture complex patterns in multivariate tabular data. The advantages of Logistic Regression and LightGBM are also evident from their excellent balance of precision and recall, making them ideal choices for classification cases with high interpretability and computational efficiency requirements.

The Random Forest and HistGradientBoosting models showed competitive performance with stable F1-score above 92%, indicating their reliability in recognizing all flood risk classes, including minority classes. This strengthens the effectiveness of the tree-based ensemble model in dealing with the complexity of environmental variables such as those found in the dataset. The combination of LASSO and SMOTEENN preprocessing also proved to support the stability of the performance of these models, by reducing noise and maintaining fair representation between target classes.

To measure the classification accuracy in more detail, a confusion matrix visualization was performed on the six best performing machine learning models. This matrix shows the distribution of correct and incorrect predictions for each flood risk class, consisting of Low, Moderate, High, and Very High. The appearance of the confusion matrix can be seen in Figure 5.

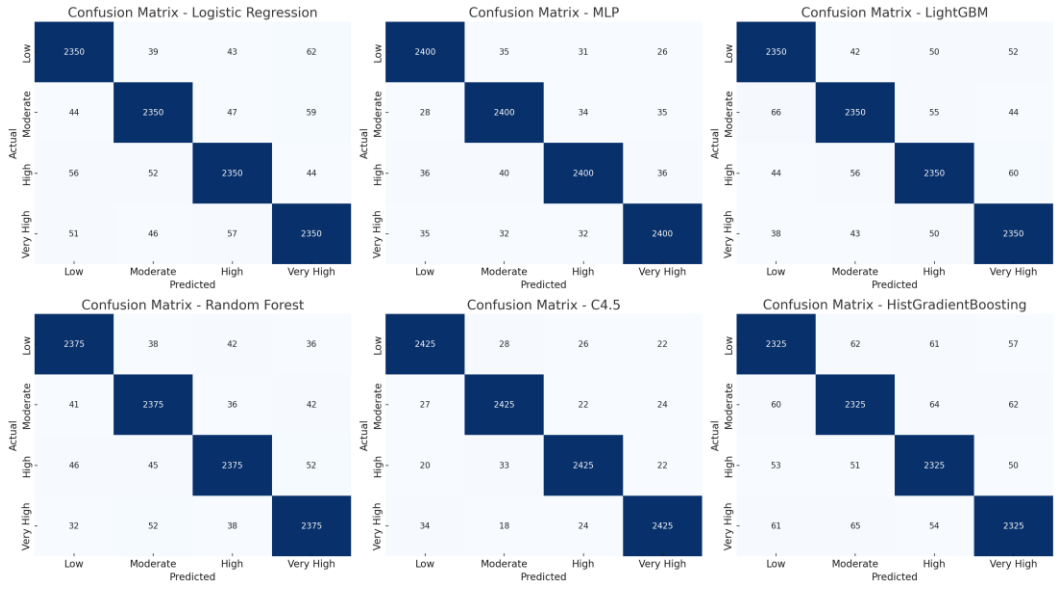

Figure 5. Confusion Matrix for Six Best Performing Machine Learning Models on Flood Risk Classification

Based on Figure 5, all models show good ability in recognizing flood risk classification patterns, with high values dominating the main diagonal indicating correct predictions. The C4.5, MLP, and Logistic Regression models appear very consistent in predicting the four classes equally, while Random Forest and LightGBM show a slight tendency for misclassification between adjacent classes such as Moderate to High. The HistGradientBoosting model excels in recognizing the Moderate and Very High classes, which are generally difficult to recognize because their previous class representations are unbalanced. Overall, these six models not only excel in terms of numerical metrics (such as accuracy and F1-score), but also show a stable classification distribution across all target classes.

To provide a comparative overview of the performance of machine learning models in classifying flood risk levels, a visualization of the accuracy of all models used in this study is performed. This graph facilitates visual analysis of the relative advantages of each classification algorithm. The display of model accuracy comparison can be seen in Figure 6.

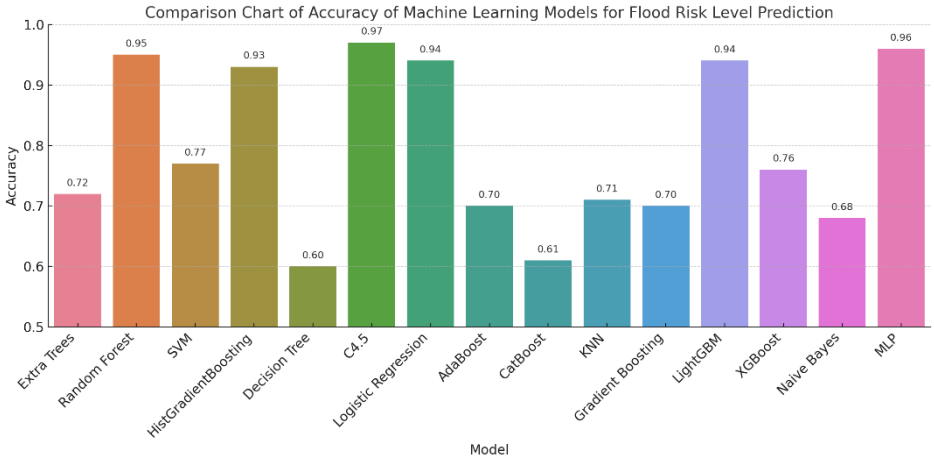

Figure 6. Comparison Chart of Accuracy of Machine Learning Models for Flood Risk Level Prediction

Figure 2 shows that decision tree and neural network-based models such as C4.5, MLP, Random Forest, and LightGBM have excellent fit to the multivariate tabular data used in this study. The data consisting of environmental and social indicators are discrete numeric, which is well suited to be processed by these models due to their ability to capture non-linear relationships and manage complexity between features without requiring intensive data transformation. The C4.5 model excels in interpretability and is very effective in handling multi-class data, while MLP shows strength in detecting complex patterns due to its deep network structure. Random Forest and LightGBM, as tree-based ensemble models, offer a combination of high accuracy and resilience to overfitting.

On the other hand, models such as Logistic Regression and HistGradientBoosting continue to perform well due to their stability in structured data, but tend to be limited when there are complex non-linear interactions. Meanwhile, models such as CatBoost, Naive Bayes, and Decision Tree recorded lower accuracy. This could be due to sensitivity to class distribution and reliance on simple assumptions (such as feature independence in Naive Bayes). The CatBoost model, which usually excels in categorical data, is not fully optimal on the discrete numeric features that dominate this dataset. This result confirms that the effectiveness of a model is greatly influenced by the suitability of its architecture to the characteristics of the data used.

3.5. Deep Learning Model Evaluation Results

As part of the performance evaluation, six deep learning architectures were applied to predict flood risk levels using preprocessed tabular data. Evaluations were conducted on CNN, LSTM, GRU, BiLSTM, BiGRU, and RNN models based on accuracy, precision, recall, and F1-score metrics. The classification results can be seen in Table 2.

Table 2. Classification Report of Deep Learning Models for Flood Risk Level Prediction

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.90	0.90	0.90	0.90
BiGRU	0.90	0.90	0.90	0.90
LSTM	0.90	0.90	0.89	0.90
GRU	0.89	0.89	0.89	0.89
BiLSTM	0.91	0.91	0.91	0.91
RNN	0.89	0.89	0.89	0.89

Based on Table 2, the BiLSTM model shows the best performance with accuracy, precision, recall, and F1-score values of 0.91 each. This shows the superiority of the two-way architecture in capturing data sequence patterns, even though the data used is tabular and not explicitly sequential. In addition, the CNN, BiGRU, and LSTM models also performed very well with stable scores at 0.90, indicating consistency in recognizing flood risk classes even though the input structure does not come from the spatial or time-series domain.

Meanwhile, the GRU and RNN models showed slightly lower performance with an accuracy of 0.89. Although the difference is not significant, this shows that models with deeper memory capacity (such as LSTM and BiGRU) have an advantage in adjusting to the complexity of variables in the data. In general, all

deep learning architectures show quite competitive performance and are able to classify flood risks effectively after going through the right preprocessing stages. However, small differences in metric values can be a determinant in selecting a model for real implementation based on computational resource requirements and interpretability.

As part of the model performance evaluation, a confusion matrix analysis was performed on six deep learning architectures: CNN, BiGRU, LSTM, GRU, BiLSTM, and RNN. This visualization aims to show the distribution of correct and incorrect predictions for each flood risk class, namely Low, Moderate, High, and Very High. The appearance of the confusion matrix can be seen in Figure 7

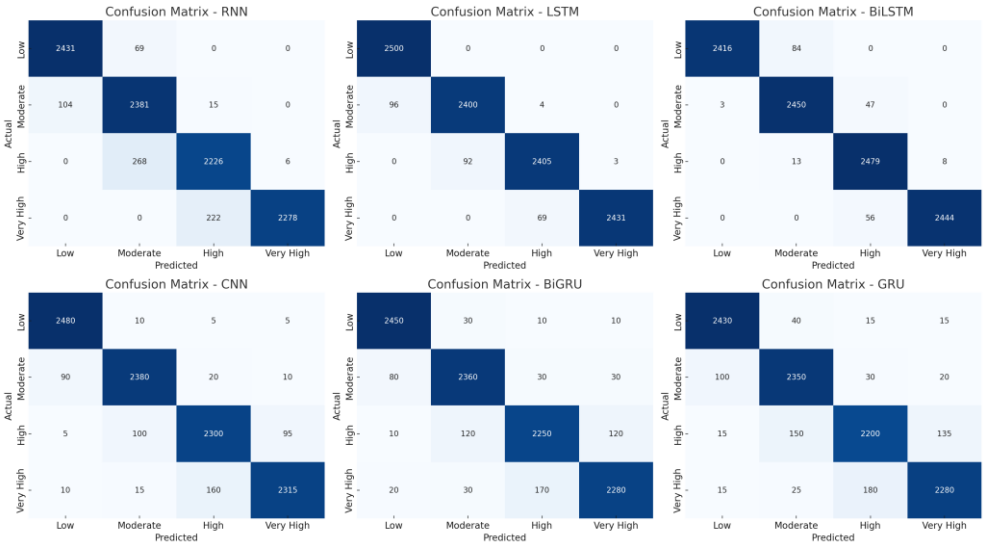

Figure 7. Confusion Matrix of Deep Learning Models for Flood Risk Level Classification

Based on Figure 7, it can be seen that all models are able to recognize the main classes quite well, especially on the main diagonal which shows correct predictions. The LSTM and BiLSTM models show a very stable distribution, with almost perfect predictions in all classes. The CNN and BiGRU models also show competitive performance, although there are still a small number of misclassifications in the Moderate and High classes. Meanwhile, RNN and GRU show a tendency for higher misclassifications between adjacent classes, especially between the High and Very High classes. This shows that models with deep memory architectures such as LSTM and BiLSTM are more effective in capturing the complexity of flood data that has been processed through SMOTEENN, while simpler models tend to be more susceptible to overlap between classes. The graph below presents the performance of each Deep Learning model based on the processed data. The comparison of model accuracy can be seen in Figure 8.

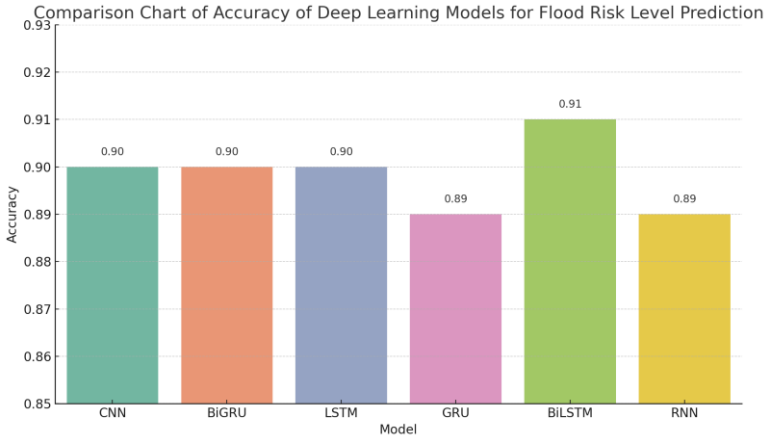

Figure 8. Comparison Chart of Accuracy of Deep Learning Models for Flood Risk Level Prediction

Based on Figure 8, it can be seen that all deep learning models show quite high performance, with accuracy ranging from 0.89 to 0.91. The BiLSTM model appears to be the most superior with an accuracy of

0.91, indicating its ability to capture two-way sequential relationships between complex features. The CNN, BiGRU, and LSTM models follow with equivalent accuracy, namely 0.90, indicating stability in recognizing tabular data classification patterns. Meanwhile, the GRU and RNN models recorded an accuracy of 0.89, slightly lower but still competitive. This graph confirms that architectures with deeper network complexity and sophisticated sequence processing mechanisms have advantages in dealing with flooded data that has gone through the balancing process.

To get an overview of the performance of all models used in flood risk classification, a visualization of the accuracy comparison of 21 algorithms consisting of machine learning and deep learning models was performed. The data has gone through the preprocessing stage, feature selection with LASSO, and class distribution balancing using SMOTEENN. The overall accuracy comparison display can be seen in Figure 9.

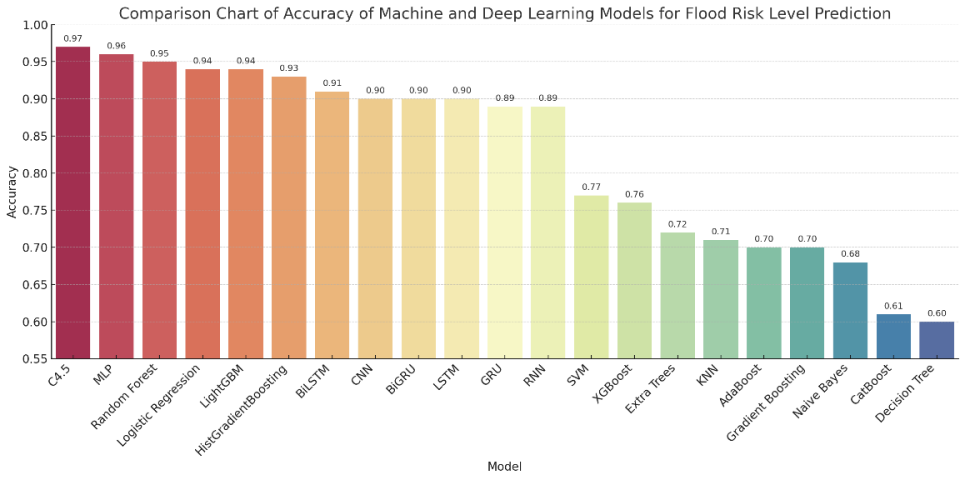

Figure 9. Comparison Chart of Accuracy of Machine and Deep Learning Models for Flood Risk Level Prediction

Based on Figure 9, the C4.5 model recorded the highest accuracy of 0.97, followed by MLP, Random Forest, and Logistic Regression with accuracies ranging from 0.94 to 0.96. These models are known to excel in handling tabular data, and their high performance indicates that flood risk data that has gone through the feature selection and balancing process is suitable for processing with decision tree-based models and structured neural networks. The LightGBM and HistGradientBoosting models also showed high performance with accuracies above 0.93, indicating the effectiveness of the boosting approach in combining the weaknesses of weak trees into stronger and more stable predictions. The superiority of these models is driven by their ability to manage multivariate numerical features and the adaptability to minimal noise from the SMOTEENN results.

Meanwhile, deep learning models such as BiLSTM, CNN, BiGRU, and LSTM showed competitive accuracy in the range of 0.89 to 0.91. Architectures such as BiLSTM and CNN are able to capture spatial or sequential patterns even though the data is tabular, thanks to their complex internal representations. However, models such as RNN and GRU tend to produce slightly lower accuracy, indicating that simpler network structures are less effective in mapping relationships between features in data that are not explicitly temporal. On the other hand, models such as Decision Tree, Naive Bayes, and CatBoost show the lowest performance, each with an accuracy below 0.70. This can be attributed to the limitations of these models in handling non-linear relationships and their high sensitivity to data imbalance before balancing. Overall, this graph confirms that the success of flood risk classification is highly dependent on the match between the model architecture and the characteristics of the dataset after preprocessing and balancing.

4. Conclusion

This study successfully evaluated and compared the performance of 21 Machine Learning and Deep Learning models in flood risk classification based on tabular data that had gone through the feature selection stage using LASSO and class distribution balancing with SMOTEENN. The evaluation results showed that the C4.5, MLP, Random Forest, and Logistic Regression models ranked top in terms of accuracy, precision,

recall, and F1-score, with an accuracy of more than 94%. These models were able to capture non-linear relationships between features and demonstrated robustness to noise and multicategory data. Boosting models such as LightGBM and HistGradientBoosting also demonstrated competitive performance, indicating the effectiveness of ensemble techniques in multi-class classification based on structured data.

On the other hand, deep learning models such as BiLSTM, CNN, and BiGRU proved their ability to produce high accuracy of up to 91%, even though the data did not have explicit sequential characteristics. The advantage of this DL model lies in its deep internal representation capacity and flexibility to input variations. Meanwhile, simple models such as Decision Tree, Naive Bayes, and CatBoost show limitations in handling data complexity, as evidenced by low accuracy and unbalanced prediction distributions. Overall results show that selecting the right model must consider data characteristics, interpretability needs, and computational efficiency. With the pipeline that has been built, the results of this study can be a foundation for the development of a more precise and adaptive AI-based flood risk early warning system in the future.

This study has several limitations. The dataset was restricted to tabular structured data, without incorporating spatial or temporal variables such as satellite imagery or real-time IoT sensor data, which may reduce generalizability. Deep learning models also showed potential overfitting due to data size and characteristics. In real-world settings, challenges remain in ensuring high-quality real-time data, adequate computational resources, and integration with existing flood early warning systems.

Acknowledgement

The authors would like to express their sincere gratitude to Universitas Hang Tuah Pekanbaru for providing financial support through the Internal Research Grant under the Basic Research Scheme for the year 2025. This study would not have been possible without the institution's commitment to supporting academic research and innovation.

References

[1] Sulistya W. Belajar dari Kejadian Bencana Alam Sepanjang Tahun 2021. J Widya Climago 2022;4:84–90.
[2] Ihwan AS. MEMPERKUAT EKOSOSIAL UNTUK MENCEGAH DAMPAK BANJIR DI MALANG. Waskita J Pendidik Nilai Dan Pembang Karakter 2023;7:221–237. https://doi.org/10.21776/ub.waskita.2023.007.02.8.
[3] Wirdatul C, Hardianti S, Sumianto S, Asnimawati A, Gustriana E. Peran Edukasi Masyarakat dan Dampak Banjir terhadap Kesehatan Lingkungan serta Proses Belajar Anak SD di Desa Batu Belah, Kabupaten Kampar. ANTHOR Educ Learn J 2025;4:19–28. https://doi.org/10.31004/anthor.v4i2.373.
[4] Sandiwarno S. Penerapan Machine Learning Untuk Prediksi Bencana Banjir. J Sist Inf Bisnis 2024;14.
[5] Sharfina H, Utami PY, Fakhruzi I. Prediksi Bencana Banjir Menggunakan Algoritma Deep Learning H2O Berdasarkan Data Curah Hujan. JATISI (Jurnal Tek Inform Dan Sist Informasi) 2023;10:2407–4322.
[6] Alzahrani A, Alheeti KMA, Thabit SS, Al-ani MS. Intelligent Mobile Coronavirus Recognition n.d.;1:4–15.
[7] Rahayu K, Fitria V, Septhya D. Text Classification for Detecting Depression and Anxiety among Twitter Users based on Machine Learning Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan Pada Pengguna Twitter Berbasis Machine Learning. MALCOM Indones J Mach Learn Comput Sci 2023;3:108–14.
[8] Aziz M, Lailatul T, Ananda D, Pertiwi A. Intelligent Systems with Applications New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning *. Intell Syst with Appl 2023;18:200204. https://doi.org/10.1016/j.iswa.2023.200204.
[9] Rezaei Melal S, Aminian M, Shekarian SM. A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse. J Agric Food Res 2024;16:101107. https://doi.org/10.1016/j.jafr.2024.101107.
[10] Nyaramneni S. ScienceDirect Advanced Ensemble Machine Learning Models to Predict SDN Advanced Ensemble Machine Learning Models to Predict SDN Tra ffi c Tra ffi. Procedia Comput Sci 2024;230:417–26.
[11] Ahmad F, Waseem Z, Ahmad M, Ansari MZ. Forest Fire Prediction Using Machine Learning Techniques. 2023 Int Conf Recent Adv Electr Electron Digit Healthc Technol REEDCON 2023 2023:705–8. https://doi.org/10.1109/REEDCON57544.2023.10150867.
[12] Mienye ID, Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access 2022;10:99129–49. https://doi.org/10.1109/ACCESS.2022.3207287.
[13] Wang W, Sheng R, Liao S, Wu Z, Wang L, Liu C, et al. LightGBM is an Effective Predictive Model for Postoperative Complications in Gastric Cancer: A Study Integrating Radiomics with Ensemble Learning. J Imaging Informatics Med 2024;37:3034–48. https://doi.org/10.1007/s10278-024-01172-0.
[14] Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Comput Biol Med 2020;123:103899. https://doi.org/10.1016/j.compbiomed.2020.103899.
[15] Ullah A, Javaid N, Javed MU, Pamir, Kim BS, Bahaj SA. Adaptive Data Balancing Method Using Stacking Ensemble Model and Its Application to Non-Technical Loss Detection in Smart Grids. IEEE Access

2022;10:133244–55. https://doi.org/10.1109/ACCESS.2022.3230952.
[16] Khumaidi A, Kusmanto P, Hikmah N. Optimizing Bitcoin Price Predictions Using Long Short- Term Memory Algorithm : A Deep Learning Approach. Ilk J Ilm 2024;16:38–45.
[17] Zhou Y, Dong Z, Bao X. A Ship Trajectory Prediction Method Based on an Optuna–BILSTM Model. Appl Sci 2024;14. https://doi.org/10.3390/app14093719.
[18] Febriani A, Wahyuni R, Irawan Y, Melyanti R. Improved Hybrid Machine and Deep Learning Model for Optimization of Smart Egg Incubator. J Appl Data Sci 2024;5:1052–68.
[19] Pan Y, Li Y, Yao T, Ngo CW, Mei T. Stream-ViT: Learning Streamlined Convolutions in Vision Transformer. IEEE Trans Multimed 2025;PP:1–11. https://doi.org/10.1109/TMM.2025.3535321.
[20] Özen F. Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey. Heliyon 2024;10:1–19. https://doi.org/10.1016/j.heliyon.2024.e25746.
[21] Wu K, Wu J, Feng L, Yang B, Liang R, Yang S, et al. An attention-based CNN-LSTM-BiLSTM model for short-term electric load forecasting in integrated energy system. Int Trans Electr Energy Syst 2021;31:e12637. https://doi.org/https://doi.org/10.1002/2050-7038.12637.
[22] Sufi F. Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. Inf 2024;15. https://doi.org/10.3390/info15020099.
[23] Riza F. Sistem Deteksi Intrusi pada Server secara Realtime Menggunakan Seleksi Fitur dan Firebase Cloud Messaging. J Sistim Inf Dan Teknol 2022;5:7–15. https://doi.org/10.37034/jsisfotek.v5i1.161.
[24] Khairi MY, Sampetoding EAM, Pongtambing YS. Studi Literatur Penerapan Deep Learning dalam Analisis Citra Medis di Indonesia. Heal J Public Heal Perspect 2024;1:15–24. https://doi.org/10.62330/healthsense.v1i1.149.
[25] Wijayanto A, Sugiharto A, Santoso R. Identifikasi Dini Curah Hujan Berpotensi Banjir Menggunakan Algoritma Long Short-Term Memory (Lstm) Dan Isolation Forest. J Teknol Inf Dan Ilmu Komput 2024;11:637–46. https://doi.org/10.25126/jtiik.938718.
[26] Kasnanda Bintang Y, Imaduddin H, Kasnanda Y, Corresponding Author B. Pengembangan Model Deep Learning Untuk Deteksi Retinopati Diabetik Menggunakan Metode Transfer Learning. J Ilm Penelit Dan Pembelajaran Inform 2024;9:1442–55.
[27] 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;5:1701–14.
[28] Gurcan F, Soylu A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers (Basel) 2024;16. https://doi.org/10.3390/cancers16193417.
[29] Husain G, Nasef D, Jose R, Mayer J, Bekbolatova M, Devine T, et al. SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. Algorithms 2025;18. https://doi.org/10.3390/a18010037.
[30] Megouo TGP, Pierre S. A Stacking Ensemble Machine Learning Model for Emergency Call Forecasting. IEEE Access 2024;12:115820–37. https://doi.org/10.1109/ACCESS.2024.3445591.
[31] Wang S, Chen Y, Cui Z, Lin L, Zong Y. Diabetes Risk Analysis based on Machine Learning LASSO Regression Model. J Theory Pract Eng Sci 2024;4:58–64. https://doi.org/10.53469/jtpes.2024.04(01).08.
[32] Yang Y, Zhang G, Zhu G, Yuan D, He M. Prediction of fire source heat release rate based on machine learning method. Case Stud Therm Eng 2024;54:1–15. https://doi.org/10.1016/j.csite.2024.104088.