Natural Disaster Classification Using MobileNet With Transfer Learning

Manuscript received November 22, 2025; revised January 2, 2026

Asri Mulyani*
Department of Computer Science
Institut Teknologi Garut
Garut, Indonesia
asrimulyani@itg.ac.id

Dede Kurniadi
Department of Computer Science
Institut Teknologi Garut
Garut, Indonesia
dede.kurniadi@itg.ac.id

Gina Suciyana
Department of Computer Science
Institut Teknologi Garut
Garut, Indonesia
2106066@itg.ac.id

***Abstract*— Natural disasters are events caused by natural phenomena that result in significant damage and pose threats to human safety. According to EM-DAT data (2000–2025), over 10,000 disasters have occurred globally, causing millions of casualties and trillions of US dollars in losses, including US$320 billion in 2024 due to extreme weather. These conditions underscore the importance of accurate disaster classification systems for rapid mitigation and response. This study aims to develop a natural disaster image classification model using Convolutional Neural Networks (CNN) with a Transfer Learning approach based on the efficient and lightweight MobileNetV2 architecture. The study employs the SEMMA (Sample, Explore, Modify, Model, Assess) framework as a systematic approach to organize the research workflow. The process begins with sampling, which involves collecting image data from various open sources such as Kaggle and prior literature, followed by selection, cleaning, and normalization. Data exploration is conducted to understand class distribution and identify imbalances. To address these issues, Synthetic Minority Over-sampling Technique (SMOTE) and image data augmentation (e.g., rotation, flipping, and contrast adjustment) are applied to enrich the variation of training data. The pre-trained MobileNetV2 model is then fine-tuned using the modified dataset, with hyperparameter adjustments to achieve optimal performance. Evaluation is conducted using multiple metrics, including accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC curves. Results indicate that the combination of Transfer Learning, data augmentation, and SMOTE improves model performance, achieving up to 99% accuracy and an AUC-ROC above 0.99 on public test data. Furthermore, testing on 26 private test images shows 92.31% accuracy, with 24 out of 26 images correctly classified. These findings demonstrate that the combination of MobileNetV2, augmentation, and SMOTE effectively enhances multi-class classification performance for natural disaster images. Nevertheless, these results remain empirical and require further validation. In addition, the use of the lightweight MobileNetV2 makes this approach potentially applicable to disaster mitigation support systems with limited computational resources.**

Keywords—MobilenetV2, Natural Disasters, SEMMA, SMOTE, Transfer Learning.

I. INTRODUCTION

Natural disasters are events that occur as a result of natural phenomena, which have the potential to cause significant damage to the environment and property, as well as pose a threat to human safety [1]. Law No. 24 of 2007 classifies natural disasters as including earthquakes, landslides, floods, droughts, cyclones, and volcanic eruptions [2]. Globally, the intensity of disasters has shown an increasing trend; throughout 2024, losses totaled USD 320 billion, with 93% of incidents related to extreme weather, resulting in approximately 11,000 fatalities [3]. Between 2000 and 2025, more than 10,000 disaster events resulted in the deaths of 1.6 million people and caused economic losses of USD 3.76 trillion [4]. This situation highlights the importance of a rapid and accurate disaster classification system as part of mitigation and emergency response efforts.

Advances in computer vision technology enable the use of images as a basis for automatic identification of various types of disasters. Convolutional Neural Networks (CNNs) have proven to be an effective method for image classification tasks [5]. For example, research by Anindyo Abhinowo et al. reported an accuracy of 91.62% with an F1-score, Recall, and Precision of 92% each in natural disaster classification [6]. Furthermore, Sheth et al. developed an ensemble model that achieved a higher accuracy of 92.79%, surpassing the performance of a single CNN (87.84%) with precision, recall, and F1-score values of 0.94, 0.94, and 0.93, respectively [7]. In a separate study, Aamir et al. demonstrated the superior performance of multilayered deep CNNs, which were able to identify disaster intensity with an accuracy of 99.92%, a sensitivity of 97.54%, a specificity of 98.22%, a precision of 97.79%, and an F1-score of 97.97% [8].

In addition, Cinar compared various algorithms, including KNN, SVM, Random Forest, Neural Network, and Logistic Regression, and found that the Neural Network provided the best results, with an accuracy of 90.2% and an F1-score of 0.901 [9]. A study by Sobhana et al. also confirmed that the use of data augmentation in CNN can significantly improve model accuracy, from 88% to 92% in test data [10].

Although CNNs have proven to be effective, training a CNN from scratch requires a large and balanced dataset, which is challenging in the domain of natural disasters, as these generally have a limited number of images and an uneven class distribution. Transfer learning is an effective solution because it leverages pre-trained models, such as ImageNet, to overcome these limitations. Numerous studies have demonstrated the effectiveness of this method. For example, research conducted by Alam et al. shows that MobileNet retrained through transfer learning was able to achieve 99.85% accuracy in brain tumor classification. In an unbalanced diabetic retinopathy dataset, model performance improved after applying SMOTE and Voting Classifier, increasing accuracy by 1.97% and recall by 5.43% [11]. Research by Ningrum et al. also demonstrated that combining CNN and SMOTE in soybean quality classification can achieve an accuracy of 92.42% [12]. In addition, Salehi et al. reported that

 EISSN: 2961-872X ● PISSN: 2961-8924 ● DOI: 10.56741/IISTR.esl.001682

Transfer Learning increased accuracy to 95% in medical image classification while significantly reducing training time [13].

The architecture that supports high efficiency is MobileNetV2, an advancement of MobileNetV1 that introduces Inverted Residuals and Linear Bottlenecks mechanisms. This structure significantly reduces the number of parameters while maintaining competitive accuracy and performance [14]. As reported in a study by Pradnyatama et al., MobileNetV2 was able to achieve 93.20% accuracy in butterfly species classification with much higher efficiency than heavy models such as Xception [15]. In another study, Hassan et al. showed that Transfer Learning on EfficientNetB0 was able to achieve an accuracy of 99.56% in plant disease classification [16], proving that this technique is effective in dealing with limited datasets.

In addition to architectural efficiency, improvements in model performance are also greatly influenced by the quality and balance of the training data. One common challenge in disaster image classification is the imbalance in the number of samples between classes, which can cause the model to predict the majority class more often and ignore the minority class. To overcome this problem, the Synthetic Minority Over-sampling Technique (SMOTE) is employed because it can generate synthetic samples that closely resemble the characteristics of the minority class without relying on simple duplication. Previous studies have shown that SMOTE can improve model learning quality, including in limited and unbalanced dataset scenarios, allowing models to adapt better during the training process [17].

On the other hand, image augmentation is also an important strategy in automatically expanding data variation by applying transformations such as rotation, flipping, zooming, and lighting changes. Various studies have proven that image augmentation can improve model generalization capabilities, especially when the amount of original data is limited and the visual variation between images is quite high [18],[19]. By combining image augmentation and SMOTE, the model obtains a more balanced data distribution and richer visual variation, thereby increasing learning stability and making the model more robust in classifying various types of natural disasters.

Although previous studies have reported good performance in disaster image classification, this study does not claim to introduce a new method; rather, it aims to make an incremental yet targeted empirical contribution. Specifically, this study identifies several gaps that have not been systematically examined. First, most previous studies limited classification to four disaster categories, thus failing to explore the complexity of multi-class classification with a more diverse number of categories. Second, previous studies generally focused on large architectures such as ResNet, Xception, or Inception, while empirical analysis of lightweight architectures such as MobileNetV2 in unbalanced multi-class disaster classification scenarios remains limited.

Third, although transfer learning, image augmentation, and SMOTE have been widely used separately, very few studies have analyzed the impact of combining the three in a structured manner, particularly in the context of CNN-extracted features and class imbalance in disaster data. Therefore, the contribution of this study lies in the systematic empirical evaluation of the effect of data imbalance handling through augmentation and SMOTE on the stability, generalization, and performance of the MobileNetV2 model in classifying six types of natural disasters.

In addition, many previous studies present deep learning modeling pipelines without an explicit framework that explains the systematic flow of data processing and model evaluation. This makes it difficult to analyze reproducibility and map the contribution of each stage in the deep learning pipeline. To ensure structured and reproducible analysis, this study adopts the SEMMA (Sample, Explore, Modify, Model, Assess) framework [20], [21], [22]. SEMMA is used as a systematic approach to organize the research workflow. Although SEMMA is traditionally applied in conventional data mining, this framework remains relevant because it provides a clear structure for the processes of data preparation, transformation, modeling, and evaluation, including when used in image-based deep learning scenarios. Thus, this study not only serves as an application study but also provides an empirical understanding of how data balancing strategies and lightweight architecture selection affect the performance of natural disaster image classification.

II. METHOD

A. Transfer learning

Transfer learning is a technique in deep learning that enables the use of models previously trained on large datasets to complete new tasks with smaller datasets. This technique is particularly useful in the field of computer vision because models can utilize features that have been learned from large datasets, such as ImageNet, thereby speeding up the training process and improving model accuracy on specific datasets [23]. There are two main approaches to transfer learning involving CNNs:

1) Feature extraction

In this approach, the initial layers of the pretrained model are used to extract features from new data. The initial layers in a deep learning model are responsible for recognizing general patterns, such as edges, shapes, and textures, which can be reused for various image classification tasks. By adding new layers tailored to specific tasks, the model can recognize new categories more quickly without having to train the entire network from scratch.

2) Fine Tuning

The fine-tuning approach not only utilizes features that have already been learned but also continues training on certain layers of the pre-trained model. Generally, the last few layers are adjusted to the new dataset while the initial layers are retained so that the general patterns remain relevant. In this way, the model can utilize information from the previous dataset while adjusting it to the unique characteristics of the new dataset.

B. SEMMA

This study uses the SEMMA method as a workflow framework to organize the entire model development process, rather than as a new learning method or algorithm. Although SEMMA originates from the realm of classical data mining, this framework is still suitable for deep learning-based research because it provides a structured flow from data sampling, exploration, preprocessing, and transformation to model building and evaluation. The SEMMA framework consists of five main stages: Sample, Explore, Modify, Model, and Assess [20]. which have been adapted to meet the specific needs of image analysis and CNN-based modeling. An illustration of the SEMMA process flow is shown in Fig. 1.

 EISSN: 2961-872X ● PISSN: 2961-8924 ● DOI: 10.56741/IISTR.esl.001682

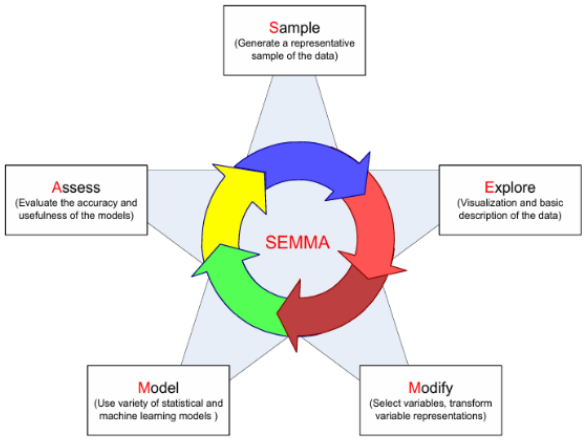

Fig. 1. SEMMA Process Flow

1) Sample

This stage is designed to select a representative subset from a large dataset, facilitating efficient processing. The resulting dataset must still capture the important patterns and trends present in the original data, ensuring its suitability for analysis and training. Additionally, careful sampling helps reduce computational complexity while maintaining the diversity and balance of the dataset categories.

2) Explore

The exploration stage involves understanding the dataset through statistical analysis and visualization techniques. During this phase, researchers identify patterns, anomalies, distributions, and key characteristics that might influence the subsequent modeling process. This process also provides insights into potential data quality issues that may affect classification accuracy.

3) Modify

In the modify stage, data transformation and adjustments are made, which can include cleaning, normalization, feature extraction, data augmentation, or the application of SMOTE. This ensures that the dataset is properly prepared for modeling. Furthermore, these modifications contribute to improving data consistency and reducing the risk of overfitting during model training.

4) Model

At this point, machine learning or deep learning algorithms are employed to build predictive models using the modified data. This stage encompasses model training and parameter optimization to enhance performance. Various evaluation metrics are also considered to determine the most effective model for natural disaster image classification.

5) Assess

The final stage assesses model performance by utilizing validation and test datasets. A confusion matrix is typically employed to evaluate the model's accuracy in classifying data and to identify any errors. This evaluation process is crucial for optimizing model performance and ensuring that the classifications produced are both reliable and accurate, following the established metrics [24].

a) Accuracy

made by the model compared to the total number of data points, providing an overall indication of the model's performance. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{n} \quad (1)$$

b) Precision

Precision assesses the proportion of true positive predictions relative to all positive predictions, reflecting how accurately the model predicts positive cases. It is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

c) Recall

Recall evaluates the model's ability to identify all relevant positive instances in the dataset, measuring the proportion of true positives among actual positives. The formula for recall is:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

d) F1-Score

The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both aspects of model performance. It is particularly useful when the class distribution is uneven. The calculation is:

$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

With :

TP	:	Positive data predicted correctly
TN	:	Negative data predicted correctly
FP	:	Type 1 error, where negative data is predicted as positive
FN	:	Type 2 error, where positive data is predicted as negative.
n	:	Total number of data/examples

III. RESULT

This study employs the SEMMA methodology to construct a natural disaster image classification model. The entire process follows the SEMMA workflow, starting with data sampling (Sample), data characteristic exploration (Explore), data modification and engineering (Modify), classification model construction (Model), and ending with model performance evaluation (Assess). The results of each stage are described as follows. Furthermore, the implementation of this framework enables systematic analysis, improves classification accuracy, supports efficient feature extraction, and provides reliable evaluation outcomes for disaster-related image datasets.

A. Sample

This stage focused on collecting and selecting relevant image data for the classification of natural disasters. The primary sources of data were public platforms, including Kaggle, as well as several previous studies. From Kaggle, 930 tropical cyclone images, 1,349 earthquake images, 1,072 flood images, and 1,078 forest fire images were obtained [25]. As a supplement, study [26] added 375 flood images, 77 landslide images, and 257 earthquake images. The volcanic eruption category was obtained from [27] with 99 images sourced from [28], while an additional 770 landslide images were obtained from the study [29]. These diverse datasets ensured broader representation of disaster characteristics, reduced data imbalance across categories, enhanced the model's generalization capability, and supported the

development of a more accurate and reliable image classification system for automatically identifying various types of natural disasters.

TABLE I. SAMPLE DATASET

Class Disaster	Sample Dataset	
Earthquake		
Forest Fire		
Flood		
Cyclone		
Land-slide		
Volcanic Eruption		

Table I illustrates the visual variations of each disaster class. To ensure data quality, data selection and cleaning processes were carried out using the following criteria:

1. The image must visually represent the disaster (the environmental impact or typical disaster phenomena, for example).
2. images with very low resolution, excessive watermarks, or irrelevant to the disaster context were eliminated, and
3. Image duplication between sources was removed based on visual inspection and file name similarity.

Label verification was performed manually through visual review of each image to ensure consistency between the label and its visual content. Although this verification was subjective and did not involve independent annotators, this step was taken to minimize labeling errors resulting from the merging of data from various sources.

The final dataset was then combined and stored in a compressed format (ZIP) on Google Drive [30] for model processing and training during the research.

B. *Explore*

The Explore stage is conducted to understand the structure and distribution of data, identifying problems in the dataset before the model training process begins. A preliminary analysis is conducted on the number of images in each class.

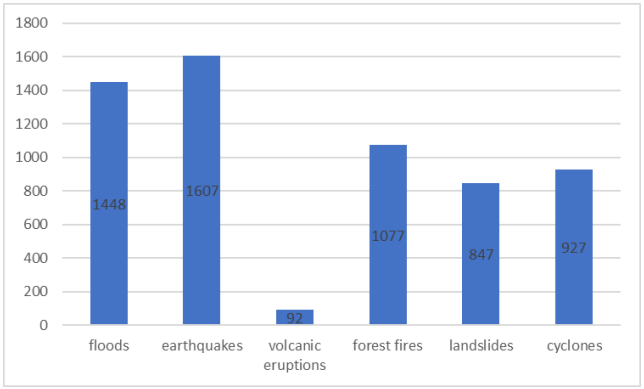

Fig. 2. Amount Of Data According To Disaster Category

Based on Fig. 2, the earthquake class has the highest number of images (1,607), followed by floods (1,448), forest fires (1,077), tropical cyclones (928), and landslides (847). The volcanic eruption class has the lowest number of images, namely 92 images. This distribution shows an imbalance between classes that needs to be addressed in the Modify stage, especially for the volcanic eruption class.

In addition to distribution, this stage also involves checking image quality, including resolution, visual diversity, and the absence of noise, such as blurred images or inappropriate internet search results. The results of this exploration form the basis for determining the augmentation strategy and the use of SMOTE to balance the sample size before the modeling process.

C. *Modify*

The Modify stage aims to enhance the quality, representation, and balance of the data prior to the model training process. The first step is to normalize the images using the preprocess_input function from the MobileNetV2 architecture, which converts pixel values from the range [0, 255] to the scale [-1, 1]. This normalization ensures that each image has a uniform distribution of values, thereby improving the stability of the training process.

Next, data augmentation is performed using ImageDataGenerator from Keras to enrich the visual variety in the dataset. The augmentation techniques used include rotation up to 30°, horizontal and vertical translation, zoom, and horizontal flip. Each image is expanded twice, so that the total amount of data increases from 5,999 original images to 17,997 images.

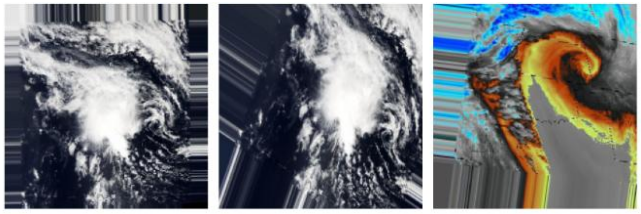

Fig. 3. Dat.aset Before Augmentation

EISSN: 2961-872X ● PISSN: 2961-8924 ● DOI: 10.56741/IISTR.esl.001682

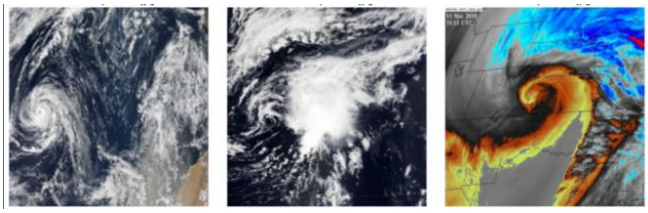

Fig. 4. Dataset After Augmentation

Fig. 3 and 4 show a comparison between the original image and the augmented image. The augmentation process produces variations in viewing angles and new visual compositions, for example, in images of floods or forest fires, allowing the model to learn more diverse visual patterns. Thus, augmentation not only increases the amount of data but also improves the model's generalization ability.

The next step is feature extraction using the pretrained ImageNet MobileNetV2 architecture with the configuration include_top=False. All augmented images are processed through this base network, then projected using Global Average Pooling, resulting in a 1,280-dimensional feature vector. This process generates a total of 17,997 feature vectors, representing the key visual characteristics of each image in a concise and informative manner.

To address class imbalance, particularly in categories with a limited number of images, such as volcanic eruptions, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) to the 1,280-dimensional features extracted by MobileNetV2. This approach aims to improve the representation of minority classes in the feature space learned by the pre-trained model.

The application of SMOTE in high-dimensional feature space is not without challenges, as the synthetic interpolation process has the potential to affect the data manifold structure and class boundaries. Therefore, SMOTE in this study is not treated as a universal solution to data imbalance, but rather as an empirical approach to evaluate its impact on classification performance.

In this study, the image augmentation process and SMOTE application were performed before dividing the data into training and validation sets.This approach allows for improved representation of minority classes and training data variation, but also has methodological consequences.

In particular, the application of SMOTE before data division has the potential to cause similarities in distribution between training and validation data due to the existence of synthetic samples in the feature space. This sequence has the potential to improve model performance, but also has methodological limitations because it can cause similarities in data distribution between the two sets, particularly due to the existence of synthetic samples in the feature space. This condition has the potential to produce optimistic performance estimates and increase the risk of indirect information leakage.

Therefore, evaluation results in scenarios involving SMOTE must be interpreted carefully and positioned primarily for comparative analysis between scenarios, not as absolute claims about the model's generalization capabilities. This approach allows for a more controlled understanding of the influence of data balancing on model stability and performance in the context of the dataset used. After the SMOTE process, the dataset is then divided into training and validation data with a ratio of 80:20 and 90:10.

D. Model

Model training in this study was conducted through three main scenarios designed to evaluate the effect of augmentation and oversampling techniques on the performance of natural disaster image classification. The first scenario utilized the original data without augmentation or oversampling techniques as a baseline, allowing model performance to be evaluated purely based on the initial dataset conditions. The second scenario utilized image augmentation techniques, specifically those implemented using the ImageDataGenerator from Keras, which included rotation up to 30°, horizontal flipping, translation, and zoom. The purpose of augmentation was to enrich the visual diversity of the dataset, allowing the model to obtain a wider variety of viewing angles, texture patterns, and image compositions, thereby improving its generalization ability.

The third scenario combines augmentation with the Synthetic Minority Over-sampling Technique (SMOTE), which is applied after the feature extraction process. This approach is used to address class imbalance that can cause model bias towards the majority class. All images, both original and augmented, were extracted using the MobileNetV2 architecture that had been pretrained on ImageNet by removing the fully connected part (include_top=False). The extraction results were 1,280-dimensional feature vectors from the GlobalAveragePooling2D layer, which were then used as input for SMOTE. This process produces a feature dataset of 28,926 samples with a more balanced class distribution.

The classification model architecture used consists of several fully connected layers of 256 and 128 units, each equipped with a dropout of 0.5 to reduce the risk of overfitting. The output layer uses softmax according to the number of target classes. The model is compiled using the Adam optimizer and the sparse_categorical_crossentropy loss function because the labels are represented as integers. The training process is conducted over 30 epochs with a batch size of 32.

Performance evaluation is carried out using two data division schemes, namely 90% training data and 10% test data, and 80% training data and 20% test data. These variations in division were used to analyze the extent to which the proportion of training data affects model performance and to comprehensively compare the results of the three scenarios. This approach provides a comprehensive overview of the most optimal training strategy in the context of natural disaster image classification with a relatively limited and unbalanced dataset.

E. Asses

The stage was conducted to evaluate the performance of the natural disaster classification model based on three different training scenarios, namely without augmentation and SMOTE (Scenario 1), using image augmentation (Scenario 2), and a combination of augmentation and SMOTE (Scenario 3). The evaluation was conducted using the accuracy (Ac), precision (Prec), recall (Rec), F1-score (F1), and AUC-ROC (AUC-ROC) metrics, with varying numbers of epochs (Ep) in each scenario. The evaluation results are presented in a table, which contains the overall performance of the model and the classification performance per disaster class. To make the data presentation more concise, each type of disaster is represented by a code: FL for floods, EQ for earthquakes, VE for volcanic eruptions, FF for forest fires, LS for landslides, and CY for cyclones. The following table presents the evaluation results for the three scenarios, including overall performance and performance by disaster class.

TABLE II. MODEL EVALUATION

SC	Split Data	Ep	Ac	AUC ROC	Class	Prec	Rec	F1
Sc1	80:20	30	0.21	0.50	FL	0.25	0.26	0.25
					EQ	0.28	0.28	0.28
					VE	0.00	0.00	0.00
					FF	0.17	0.19	0.18
					LS	0.17	0.16	0.17
					CY	0.14	0.14	0.14
Sc2	90.10	69	0.98	0.99	FL	0.96	0.97	0.97
					EQ	0.97	0.98	0.97
					VE	0.93	0.90	0.92
					FF	0.99	0.98	0.98
					LS	0.99	0.99	0.99
					CY	0.99	0.98	0.98
Sc2	80.20	100	0.97	0.99	FL	0.95	0.95	0.95
					EQ	0.95	0.97	0.96
					VE	0.94	0.92	0.93
					FF	0.98	0.96	0.97
					LS	0.99	0.98	0.98
					CY	0.99	0.99	0.99
Sc3	90.10	85	0.99	0.99	FL	0.97	0.97	0.97
					EQ	0.98	0.98	0.98
					VE	1.00	1.00	1.00
					FF	0.99	0.99	0.99
					LS	1.00	1.00	1.00
					CY	0.99	1.00	0.99
Sc3	80.20	85	0.98	0.99	FL	0.95	0.97	0.96
					EQ	0.97	0.96	0.96
					VE	1.00	1.00	1.00
					FF	0.99	0.98	0.98
					LS	0.99	0.99	0.99
					CY	0.99	1.00	0.99

The stark contrast in performance between Scenario 1 and Scenarios 2–3 shows that the baseline in Scenario 1 does not represent optimal training conditions. This scenario is designed as a basic comparison without augmentation and data balancing, so the model has the potential to experience relative under-training, especially in learning minority class representations in unbalanced datasets.

Furthermore, the baseline in Scenario 1 does not involve extensive hyperparameter tuning, so the resulting performance cannot be considered the upper limit of the MobileNetV2 architecture's capabilities on this dataset. Therefore, the large performance gap between the initial and advanced scenarios does not fully reflect the superiority of the proposed method, but is also influenced by differences in training conditions and data complexity.

The application of image augmentation in Scenario 2 contributes to an increase in training data variation, allowing the model to learn more stable visual patterns. Furthermore, the combination of augmentation and SMOTE in Scenario 3 improves the representation of minority classes, which is reflected in improved performance metrics.

Therefore, although these results provide empirical evidence that augmentation and data balancing strategies can improve model performance, the very high AUC and F1-score values cannot be directly interpreted as indicators of perfect generalization. Additional evaluation using cross-validation or external datasets is needed to confirm the stability and reliability of the model in more diverse operational scenarios.

1) Scenario 1

In Scenario 1, the model was trained without augmentation and class balancing with an 80:20 data split. The results showed very low performance, with an accuracy of 0.21 and an AUC ROC of 0.50. The precision, recall, and F1-score values per class were unstable; the flood and earthquake classes only achieved an F1-score of around 0.25–0.28, while the volcanic eruption, landslide, and tropical cyclone classes were not detected at all.

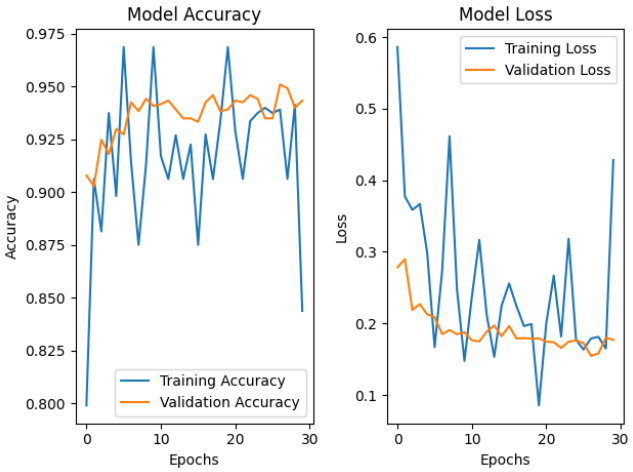

Fig. 5. Accuracy graph and loss scenario 1

The accuracy graph shows unstable training, with training accuracy increasing but fluctuating, while validation accuracy tends to stagnate. This pattern indicates a tendency toward mild overfitting. The loss graph also shows an initial decline followed by instability, indicating that the model has not been able to learn feature representations consistently and has not reached optimal convergence. In addition, the gap between training and validation performance suggests limited generalization capability, where the model performs better on training data than on unseen data. This condition may be influenced by dataset complexity, class imbalance, insufficient training samples, or suboptimal hyperparameter settings that affect the overall stability and effectiveness of the learning process.

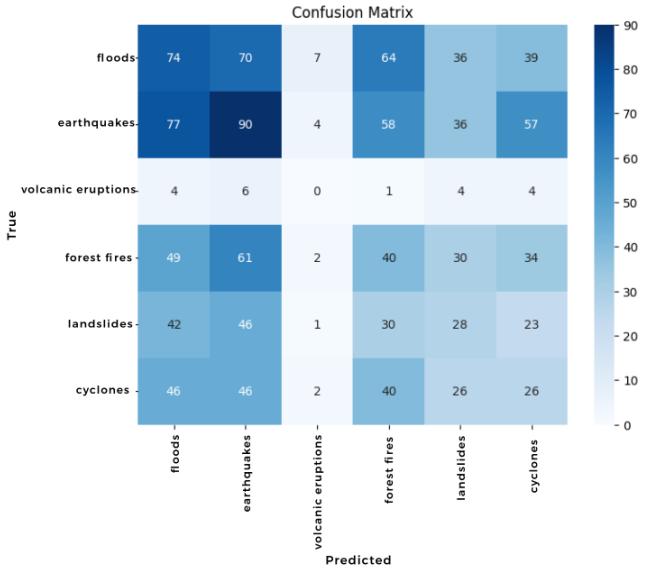

Fig. 6. Confusion Matrix For scenario 1

In addition, the confusion matrix results show a high classification error rate in several classes, particularly floods, earthquakes, and tropical cyclones, which are often confused with each other. This indicates the model's difficulty in distinguishing between classes with similar visual characteristics. Furthermore, the poor performance in the volcanic eruption class reflects the limitations of data

EISSN: 2961-872X ● PISSN: 2961-8924 ● DOI: 10.56741/IISTR.esl.001682

representation and the lack of discriminative features for that class.

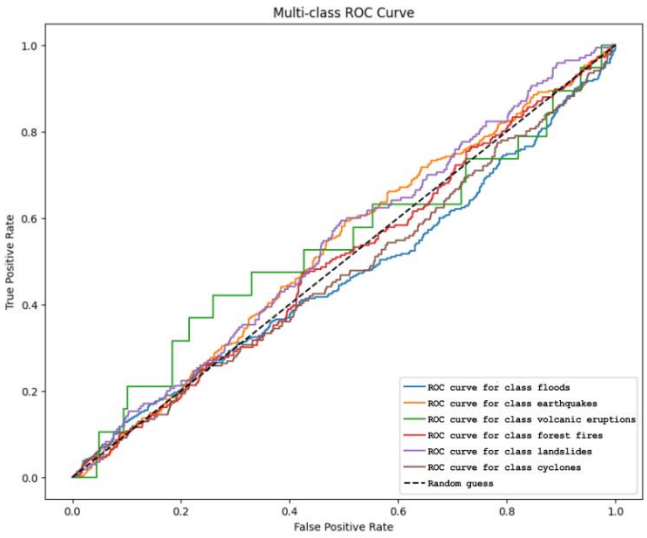

Fig. 7. multi-class roc curve scenario 1

The multi-class ROC curve shows the limitations of the model, with most curves close to the random diagonal line. An AUC value close to 0.5 indicates that the model is not yet capable of forming strong decision boundaries and still has low discriminatory power, in line with the previous accuracy and loss results. Furthermore, the overlap among class prediction probabilities suggests that the extracted features are insufficiently distinctive to separate disaster categories effectively. This outcome implies that further improvements in feature engineering, dataset quality, model architecture, or hyperparameter optimization are necessary to enhance classification sensitivity, specificity, and overall predictive reliability.

2) Scenario 2

In scenario 2, image augmentation was applied to increase the variety of training data. Testing was conducted with a 90:10 and 80:20 data split and 30–100 epochs. The results show a significant improvement in all metrics. The best performance was achieved at epoch 69 with a 90:10 split, resulting in an accuracy of 0.98 and an AUC-ROC of 0.99. All classes obtained high and stable F1-scores, with values above 0.90 in the minority class and up to 0.99 in the majority class.

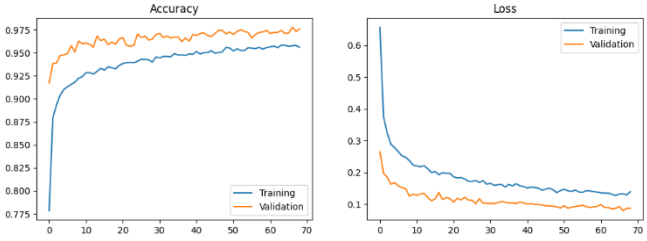

Fig. 8. Accuracy Graph and Loss Scenario 2

The accuracy and loss graphs in Scenario 2 show a stable training process. Training and validation accuracy increased consistently with a similar pattern, while the loss value decreased gradually without sharp fluctuations. The consistency of these two graphs indicates that the learning process was effective and that the model was able to extract data patterns stably at each epoch.

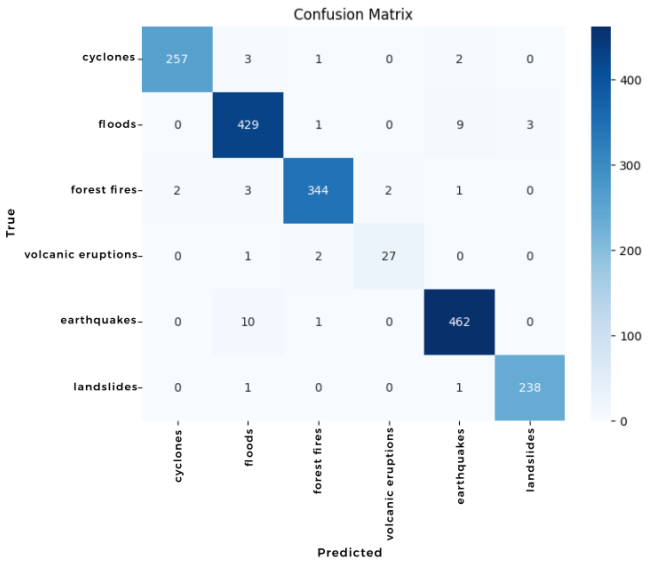

Fig. 9. Confusion Matrix For Scenario 2

The confusion matrix shows a dominance of diagonal values with minimal classification errors, indicating that the model is able to distinguish each disaster class well. Compared to the previous scenario, the misclassification rate has decreased significantly, demonstrating an improvement in the model's discriminatory ability after applying data augmentation.

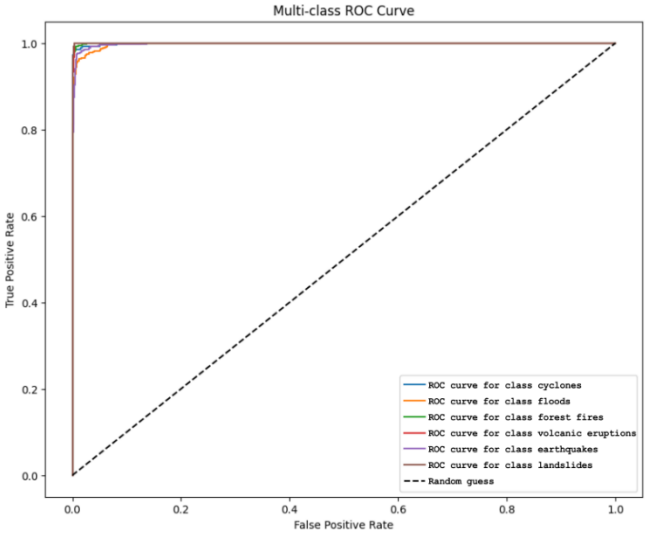

Fig. 10. Multi-class ROC curve scenario 2

The multi-class ROC curve shows an AUC close to 1.00 across all classes, indicating excellent classification capabilities. This improvement compared to the initial scenario indicates stronger model accuracy and generalization against disaster image variations.

3) Scenario 3

In this scenario, combining image augmentation and SMOTE to balance the number of samples in each class showed the best performance compared to the other two scenarios. The model achieved an accuracy and AUC ROC of 0.99 at epoch 85 with a 90:10 data split, while the F1-score per class was very high and evenly distributed, including a perfect score of 1.00 for the volcanic eruption and landslide classes. These results show that class balancing with SMOTE effectively reduces bias towards the majority class.

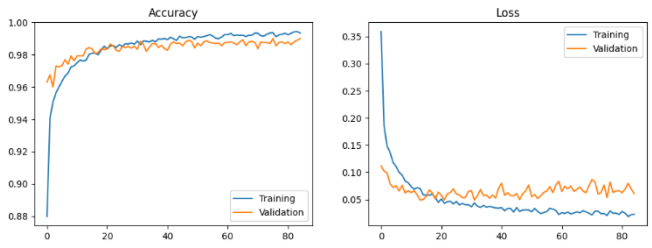

Fig. 11. Accuracy Graph And Loss Scenario 3

The accuracy and loss graphs in this scenario show a stable learning pattern and a balance between training and validation. Accuracy increases consistently, with validation accuracy in several epochs even exceeding training accuracy, indicating that the model can generalize well. The loss values in training and validation also decrease in parallel, without a significant gap, indicating no indications of overfitting or underfitting during the training process.

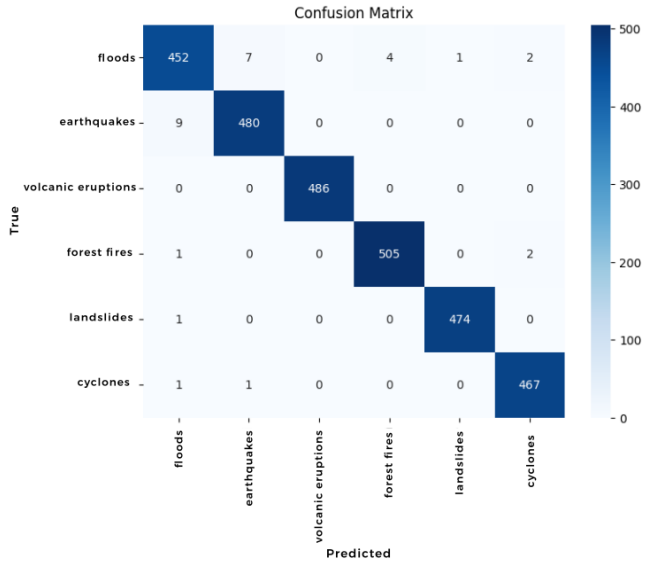

Fig. 12. Confusion matrix scenario 3

The confusion matrix in Scenario 3 shows a very high dominance of diagonal values, including perfect performance on several minority classes. These results indicate that the model has very strong classification capabilities after applying data augmentation and balancing using SMOTE.

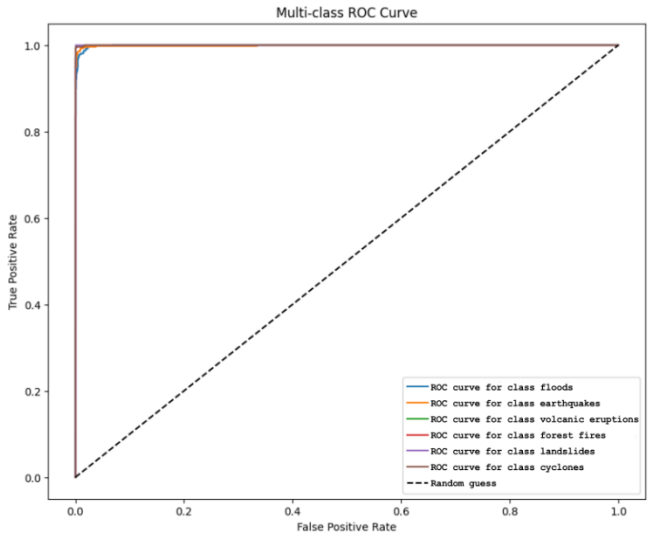

Fig. 13. Multi-class ROC curve scenario 3

The multi-class ROC curve shows that all classes have very high AUCs and are close to the ideal point in the upper left corner of the graph, indicating that the model has good sensitivity and specificity as well as strong discriminatory performance in all disaster categories. Although the AUC and confusion matrix results are nearly perfect, these findings need to be interpreted with caution. The dataset compiled from various sources and the application of SMOTE to the extracted features may cause the model to learn patterns that are highly specific to the training data, potentially leading to overfitting or performance optimism. Therefore, these results should be viewed as an indication of the effectiveness of the approach in the context of the dataset used, rather than as a guarantee of absolute generalization to natural disaster data outside the training distribution, and require additional validation through evaluation schemes such as cross-validation or cross-dataset testing.

4) Evaluation using private data

Based on the evaluation results of the three scenarios, the best model was obtained from Scenario 3, which featured a 90:10 data split. This model was then further tested using private data that had never been involved in the training or validation process. This data consisted of 26 disaster images obtained from BPBD Garut, online media reports, and other internet sources.

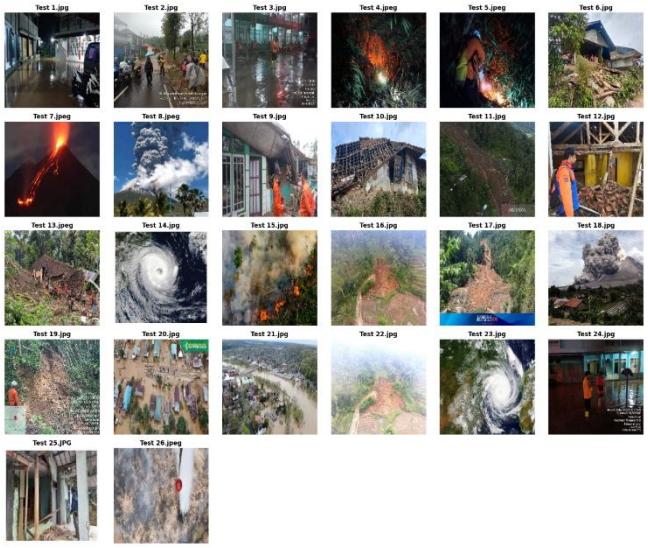

Fig. 14. Data Test

After testing with private data, the model's prediction results for all test images are summarized in the following table, which presents a comparison between the actual labels and the model's predictions for 26 disaster images.

Testing results on private data indicate that the model accurately classified 24 out of 26 images, yielding an accuracy of 92.31%. Two classification errors occurred in Test 13, where a landslide image was predicted as an earthquake, and in Test 22, where a landslide image was predicted as a forest fire. These errors were likely caused by visual similarities between disaster classes, particularly in terms of the angle of the image, soil color, vegetation, or similar-looking environmental conditions. In addition, variations in lighting, image resolution, background complexity, and overlapping visual textures may have influenced the model's ability to distinguish specific disaster characteristics accurately. Despite these limitations, the overall testing performance demonstrates that the model possesses strong classification capability and

EISSN: 2961-872X ● PISSN: 2961-8924 ● DOI: 10.56741/IISTR.esl.001682

acceptable generalization performance when evaluated using previously unseen private image data.

TABLE III. TEST DATA EVALUATION

Data Image	Actual Data	Predicted Results
Test 1	floods	floods
Test 2	floods	floods
Test 3	floods	floods
Test 4	forest fires	forest fires
Test 5	forest fires	Forest fires
Test 6	Earthquakes	Earthquakes
Test 7	Volcanic eruptions	Volcanic eruptions
Test 8	Volcanic eruptions	Volcanic eruptions
Test 9	Earthquakes	Earthquakes
Test 10	Earthquakes	Earthquakes
Test 11	Landslides	Landslides
Test 12	Earthquakes	Earthquakes
Test 13	Landslides	Earthquakes
Test 14	Cyclones	Cyclones
Test 15	Forest fires	Forest fires
Test 16	Landslides	Landslides
Test 17	Landslides	Landslides
Test 18	Volcanic eruptions	Volcanic eruptions
Test 19	Landslides	Landslides
Test 20	Floods	Floods
Test 21	Floods	Floods
Test 22	Landslides	Forest fires
Test 23	Cyclones	Cyclones
Test 24	Floods	Floods
Test 25	Earthquakes	Earthquakes
Test 26	Forest fires	Forest fires

Although these findings provide initial indications of the model's generalization ability to data outside the training distribution, the evaluation conducted on private data is still limited in terms of scale and scope. The error analysis presented is still qualitative and does not include a systematic evaluation of error patterns between classes or performance metrics per class on private data. Thus, these evaluation results are more appropriately positioned as engineering validation than as conclusive scientific evidence of generalization. Therefore, these results are not used as the main basis for model performance claims, but rather as an initial indication of its potential application to data outside the training distribution.

IV. DISCUSSION

Direct quantitative comparisons with previous studies are difficult due to differences in domain, dataset size, number of classes, level of imbalance, and evaluation protocols. Previous studies have largely focused on medical or agricultural images with fewer classes and homogeneous visual characteristics, whereas this study deals with six types of natural disasters with high visual heterogeneity and diverse data sources. Previous references are used as a methodological foundation, not as a benchmark for quantitative performance.

MobileNetV2 was chosen as a lightweight, efficient architecture, but this study did not directly compare it with other architectures, such as EfficientNet-B0 or MobileNetV3, under identical experimental conditions. Therefore, claims of MobileNetV2's superiority are limited to the model configuration, pre-processing strategy, and dataset used, with further studies using uniform protocols and benchmark datasets needed for more general conclusions. The main contribution of this research lies in the empirical evaluation of the integration of transfer learning, image augmentation, and SMOTE in imbalanced disaster classification, particularly its impact on training stability, minority class performance, and model generalization ability.

Engineering validation was achieved by testing the model on real-world private data, demonstrating the practical relevance of computer vision-based disaster classification systems. The efficiency of MobileNetV2 was evaluated through training time (2–3 seconds per epoch), confirming the stability and efficiency of training in limited computing environments. However, a complete evaluation of computational costs, including inference time, memory usage, and FLOPs, has not been conducted, limiting efficiency claims to training aspects, with more comprehensive computational analysis becoming an agenda for future research.

V. CONCLUSION

This study shows that class imbalance in training data significantly affects the performance of natural disaster image classification models, especially in recognizing minority classes. The application of data augmentation and the Synthetic Minority Over-sampling Technique (SMOTE) on features extracted from MobileNetV2 has been proven to improve the representation of minority classes and consistently improve model performance across various evaluation metrics. In the best-case scenario, the model achieved high accuracy, an AUC-ROC value above 0.99, and a good F1-score in almost all classes.

However, these evaluation results still have limitations. The testing process was carried out using a single random split of the data without applying cross-validation or testing on an external benchmark dataset. This condition limits the strength of inference regarding the stability and reliability of model performance and has the potential to cause optimistic performance estimates. In addition, classification errors were still found in classes with certain visual similarities, indicating that the model is not yet fully capable of distinguishing complex visual patterns.

Additional testing using private data resulted in an accuracy of 92.31%, indicating initial generalization capabilities. However, this achievement cannot yet be considered conclusive evidence of readiness for application in a real-world environment, given that field data generally has more visual variation, noise, and lighting conditions than research datasets.

For further research, it is recommended to apply cross-validation and utilize external benchmark datasets to improve the reliability and generalization of the results. In addition, exploring alternative data balancing methods, more adaptive augmentation techniques, and evaluating other pretrained architectures such as EfficientNet, ResNet, or Vision Transformer can provide a more comprehensive understanding of model performance. More in-depth hyperparameter tuning also has the potential to improve performance stability.

REFERENCES

[1] M. F. Al Halik and L. Septiana, "Data Analysis for Predicting Disaster-Prone Areas in West Java Using K-Means Clustering Algorithm," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 6, no. 4, pp. 856–870, 2022, doi: 10.52362/jisamar.v6i4.939.

[2] BNPB, "Definition of Disaster," bnpb.go.id. Accessed: Jan. 08, 2025. [Online]. Available: https://bnpb.go.id/definisi-bencana

[3] Munich RE, "Natural disasters 2024: Climate change is showing its claws." [Online]. Available: https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2025/natural-disaster-figures-2024.html

[4] EM-DAT, "Monitoring hazards & disasters worldwide since 1988." [Online]. Available: https://www.emdat.be/

[5] W. H. Achmad, N. Saurina, N. Chamidah, and R. Rulaningtyas, "Tuberculosis Classification Modeling Using Convolutional Neural

Network," *Pros. Semin. Implementasi Teknol. Inf. dan Komun.*, vol. 2, no. 1, pp. 9–15, 2023, doi: 10.31284/p.semtik.2023-1.3989.

[6] A. Anindyo Abhinowo, R. Rizal Isnanto, and D. Eridani, "The Best Model Selection of Convolutional Neural Network Algorithm for Natural Disaster Classification," *J. Tek. Komput.*, vol. 1, no. 4, pp. 199–208, 2023, doi: 10.14710/jtk.v1i4.37656.

[7] K. A. Sheth, R. P. Kulkarni, and G. K. Revathi, "Enhancing natural disaster image classification: an ensemble learning approach with inception and CNN models," *Geomatics, Nat. Hazards Risk*, vol. 15, no. 1, p., 2024, doi: 10.1080/19475705.2024.2407029.

[8] M. Aamir *et al.*, "Natural disasters intensity analysis and classification based on multispectral images using multi-layered deep convolutional neural network," *Sensors*, vol. 21, no. 8, 2021, doi: 10.3390/s21082648.

[9] I. Cinar, "Disaster Detection Using Machine Learning Methods with Deep Features DisasterrDetectionnUsinggMachineeLearninggMethodss withhDeeppFeatures IlkayyCINAR," no. April, 2023, [Online]. Available: https://www.researchgate.net/publication/370599940

[10] M. Sobhana, S. C. Chaparala, D. N. V. S. L. S. Indira, and K. K. Kumar, "A disaster classification application using convolutional neural network by performing data augmentation," vol. 27, no. 3, pp. 1712–1720, 2022, doi: 10.11591/ijeecs.v27.i3.pp1712-1720.

[11] M. Z. Alam, T. Roy, H. M. N. Kawsar, and I. Rimi, "Enhancing Transfer Learning for Medical Image Classification with SMOTE: A Comparative Study," no. December, pp. 20–22, 2024, [Online]. Available: http://arxiv.org/abs/2412.20235

[12] A. P. Ningrum, S. Winarno, and V. Praskatama, "Classification of Soybean Quality Using Transfer Learning Convolutional Neural Network and SMOTE," vol. 5, no. 2, pp. 155–164, 2024.

[13] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustain.*, vol. 15, no. 7, 2023, doi: 10.3390/su15075930.

[14] L. Yong, L. Ma, D. Sun, and L. Du, "Application of MobileNetV2 to waste classification," *PLoS One*, vol. 18, no. 3 March, pp. 1–16, 2023, doi: 10.1371/journal.pone.0282336.

[15] M. Pradnyatama, C. A. Sari, E. H. Rachmawanto, and H. M. M. Islam, "A Comparative Analysis of Convolutional Neural Network (CNN): MobileNetV2 and Xception for Butterfly Species Classification," *J. Masy. Inform.*, vol. 16, no. 1, pp. 69–90, 2025, doi: 10.14710/jmasif.16.1.72957.

[16] S. M. Hassan, A. K. Maji, M. Jasiński, Z. Leonowicz, and E. Jasińska, "Identification of plant-leaf diseases using cnn and transfer-learning approach," *Electron.*, vol. 10, no. 12, 2021, doi: 10.3390/electronics10121388.

[17] D. Kurniadi, F. Nuraeni, and S. M. Lestari, "Implementation of the Naïve Bayes Algorithm Using Feature Forward Selection and SMOTE to Predict the Accuracy of Undergraduate Study Periods," *J. Sist. Cerdas*, vol. 5, no. 2, pp. 63–82, 2022, doi: 10.37396/jsc.v5i2.215.

[18] Rachmat Santoso, "Data Augmentation on Metal Inscriptions for Kawi Script Detection," *J. Fasilkom*, vol. 14, no. 1, pp. 234–241, 2024, doi: 10.37859/jf.v14i1.6952.

[19] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A Comprehensive Survey of Image Augmentation Techniques for Deep Learning," *Pattern Recognit.*, vol. 137, p. 109347, 2023, doi: 10.1016/j.patcog.2023.109347.

[20] J. Dean, *Big Data, Data Mining, and Machine Learning*. Hoboken, New Jersey, USA, 2014.

[21] Q. Chen and S. Lee, "A machine learning approach to predict customer usage of a home workout platform," *Appl. Sci.*, vol. 11, no. 21, 2021, doi: 10.3390/app11219927.

[22] L. Andrade-arenas, I. Rubio-paucar, C. Yactayo-arias, and C. Yactayo-arias, "Predictive models in Alzheimer ' s disease : an evaluation based on data mining techniques," vol. 14, no. 3, pp. 2988–3002, 2024, doi: 10.11591/ijece.v14i3.pp2988-3002.

[23] M. Elgendy, *Deep Learning for Vision Systems*. Shelter Island NY: Manning Publications, 2020.

[24] S. Clara, D. Laksmi Prianto, R. Al Habsi, E. Friscila Lumbantobing, and N. Chamidah, "Feature Selection Implementation in Machine Learning Classification Algorithms to Predict Income on Adult Income Dataset," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, vol. 2, no. 1, pp. 741–747, 2021.

[25] R. Roy, "Cyclone Wildfire Flood Earthquake Database," kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/rupakroy/cyclone-wildfire-flood-earthquake-database/data

[26] F. Alam, T. Alam, M. A. Hasan, A. Hasnat, M. Imran, and F. Ofli, "MEDIC: a multi-task learning dataset for disaster image classification," *Neural Comput. Appl.*, vol. 35, no. 3, pp. 2609–2632, 2023, doi: 10.1007/s00521-022-07717-0.

[27] P. M. Del Rosso, A. Sebastianelli, D. Spiller, P. P. Mathieu, and S. L. Ullo, "On-Board Volcanic Eruption Detection through CNNs and Satellite Multispectral Imagery," pp. 1–16, 2021.

[28] Global Volcanism Program, Department of Mineral Sciences, National Museum of Natural History, and Smithsonian Institution, "Database Search - Eruptions." Accessed: Apr. 19, 2025. [Online]. Available: https://volcano.si.edu/search_eruption.cfm

[29] Y. Tynchenko *et al.*, "Landslide Assessment Classification Using Deep Neural Networks Based on Climate and Geospatial Data," *Sustain.*, vol. 16, no. 16, pp. 1–26, 2024, doi: 10.3390/su16167063.

[30] G. Suciyana, "Disaster Image Dataset for Six-Class Classification." Accessed: Apr. 26, 2025. [Online]. Available: https://drive.google.com/file/d/1E47iy2JMCF0XwItUHnJ65Faai_C35P3U/view?usp=drive_link