International Journal of Advances in Data and Information Systems
Vol. 7, No. 1, April 2026, pp. 1~11
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i1.1456

Comparasion of Weather Classification Methods on Weather Images Using GLCM Features with Random Forest and Catboost Algoritms

Muhammad Noorhafizi[1], Triando Hamonangan Saragih*[2], Muhammad Itqan Mazdadi[3], Muliadi[4], Rudy Herteno[5], Hasri Akbar Awal Rozaq[6]
[1,2,3,4,5] Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia
[6] Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara

Article Info
Article history:
Received Oct 13, 2025
Revised Nov 30, 2025
Accepted Dec 18, 2025

Keywords:
Weather Classification
Image Processing
GLCM
Random Forest
Catboost

ABSTRACT
Weather image classification is an essential process for improving automated weather information systems. However, most existing studies rely on numerical meteorological data and rarely utilize the textural characteristics embedded in atmospheric imagery. This study addresses that limitation by applying the Gray Level Co-Occurrence Matrix (GLCM) for texture feature extraction combined with Random Forest (RF) and CatBoost algorithms for classification. The dataset, obtained from Kaggle, consists of 1,125 weather images categorized into four classes: cloudy, rain, shine, and sunrise. All images were uniformly normalized and augmented using four rotation angles (0°, 45°, 90°, 135°). GLCM features were extracted with a pixel distance of 1 and gray-level quantization of 8, generating four statistical attributes: contrast, correlation, energy, and homogeneity. Both algorithms were optimized through parameter tuning and evaluated using a 5-fold cross-validation scheme with an 80:20 split ratio. Results show that the Random Forest model (n_estimators = 100, max_depth = 10, random_state = 42) achieved the highest accuracy of 92.43% (±1.12), precision of 92.50%, recall of 92.43%, and F1-score of 92.42%. In comparison, CatBoost (iterations = 100, learning_rate = 0.1, depth = 6) achieved an accuracy of 68.88% (±2.31). The findings demonstrate that GLCM feature extraction combined with Random Forest offers superior stability and accuracy for weather image classification, providing a foundation for efficient and interpretable weather information systems.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Triando Hamonangan Saragih,
Faculty of Mathematics and Natural Science,
Department of Computer Science,
Lambung Mangkurat University,
Kalimantan, Indonesia
Email: triando.saragih@ulm.ac.id

1. INTRODUCTION

The development of information technology has brought about significant changes in various aspects of life, including weather information systems. Weather information systems are essentially a way to maximize monitoring of atmospheric conditions so that they can be used for evaluation, classification, and weather prediction [1]. Weather prediction is highly difficult due to the many factors that influence it, such as wind speed, rainfall, temperature, air pressure, and other atmospheric phenomena [2].

Humans can generally only observe air conditions for relatively short periods of time, such as simply sensing the temperature or observing changes in cloud cover. However, weather

conditions are far more complex. More detailed information is essential to anticipate the adverse impacts of extreme weather, such as heavy rain that can disrupt community activities. In this regard, methods based on image processing technology are beginning to be used as a new approach to weather prediction [3].

One frequently used image processing technique is the Gray Level Co-Occurrence Matrix (GLCM). This method allows for feature extraction from pixel matrix values in an image that have a specific relationship, thus forming angular or texture patterns [4]. These patterns can serve as visual representations of certain atmospheric conditions [5]. The use of GLCM on weather imagery makes weather information systems faster in identifying symptoms related to weather changes.

Machine learning algorithms, particularly those based on decision trees, are widely used for image classification due to their robustness and interpretability [6]. Among them, Random Forest (RF) and CatBoost have shown competitive performance across domains [7]. RF employs an ensemble of decision trees to minimize overfitting and improve stability, while CatBoost leverages gradient boosting to refine classification accuracy [8]. However, their comparative performance for texture-based weather imagery remains underexplored [9].

This study focuses on evaluating and comparing the performance of Random Forest and CatBoost classifiers in weather image classification using GLCM feature extraction [10]. Specifically, it investigates how differences in learning strategies—bagging in RF and boosting in CatBoost—affect classification accuracy and model stability [11].

One study on a rice leaf disease dataset with 200 test data showed that Random Forest had an optimal accuracy of 65%, recall of 65%, precision of 64%, and F-measure of 65% [12]. Another study using a monkeypox disease image dataset of 3,192 images proved that Random Forest could achieve an accuracy of 77% with the parameter n_estimator = 100 and a data ratio of 90:10. These results confirm that RF has quite good capabilities for image-based classification, although the level of accuracy can still be improved.

Based on various previous studies, the use of the GLCM feature extraction method combined with the RF and CatBoost algorithms is highly relevant for further development in weather classification. A comparison of the two is expected to provide a comprehensive overview of the effectiveness of decision tree algorithms in the context of image-based weather information systems.

This research contributes in two key aspects. Methodologically, it provides empirical evidence on the comparative performance of RF and CatBoost for texture-based weather image classification [13]. Practically, it supports the development of more accurate and accessible image-based weather information systems that can enhance early warning and decision-making processes for the public, industry, and government sectors.

2. RESEARCH METHOD

This study uses the Gray Level Co-Occurrence Matrix (GLCM) as a feature extraction method for weather imagery. The primary objective of using GLCM is to identify values that represent texture in the image so that they can be utilized in the classification process. After the feature extraction stage is carried out, this study compares two classification algorithms, namely Random Forest and CatBoost, to assess which algorithm is more optimal in classifying weather images. These two algorithms were chosen because they are known to have good performance in image-based classification problems, especially those involving texture and pattern variables. To provide a clear picture, the steps of this study are arranged in the form of a research flowchart as shown.

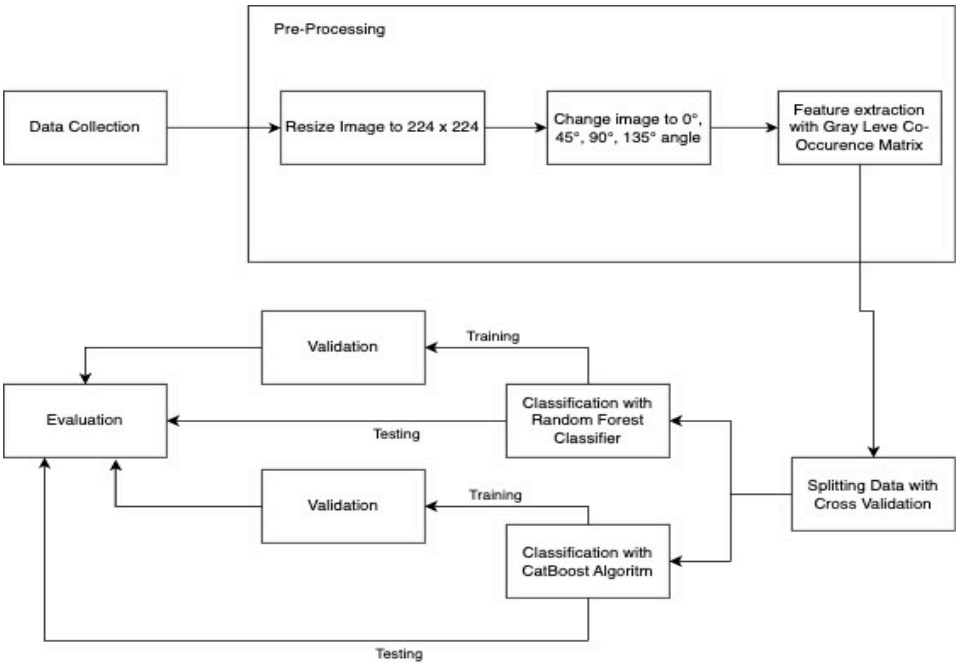


Figure 1. Flow of Research

2.1 Data Collection

Data collection is the earliest and most crucial stage in machine learning-based research. In this study, weather image data was obtained from the Kaggle website, specifically through the link "https://www.kaggle.com/datasets/pratik2901/multiclass-weather-dataset". This dataset consists of four types of weather images grouped into different folders: 357 sunrise images, 253 shine images, 300 cloudy images, and 215 rain images. This dataset is representative enough to train and test classification models.



Figure 2. Sunrise



Figure 3. Shine



Figure 4. Cloudy



Figure 5. Rain

The dataset used mostly has a resolution between 224 × 224 and 512 × 512 pixels. Based on the resolution analysis, approximately 65% of the images have a minimum size of 300 × 300 pixels, making them sufficiently detailed for texture analysis using GLCM. However, there are also

some images with smaller sizes that require prior normalization for consistency. This process is important because non-uniform image sizes can affect feature extraction results. On the other hand, the moderate variation between weather types in this dataset presents its own challenges, because the greater the variation, the more complex the process of consistent feature extraction.

2.2 Data Pre-processing

The data preprocessing stage is a crucial step to ensure that the image to be processed is in optimal condition. At this stage, several steps are taken, from adjusting the image size to data augmentation to enrich variations. The augmentation process is carried out by rotating the image at angles of 0°, 45°, 90°, and 135°. The goal is to enable the model to recognize texture patterns in various orientations, thereby improving the model's generalization ability in recognizing weather types. To ensure uniform input representation, all images were normalized using min-max scaling and augmented via rotation at 0°, 45°, 90°, and 135°. This augmentation enhances generalization by exposing models to texture variations. Normalization and augmentation were applied uniformly across all classes.

Additionally, feature extraction is performed using GLCM, which utilizes the correspondence between pixels in the image. GLCM generates a new matrix containing the frequency values of patterns between pixels. Thus, the textural characteristics of the image can be represented numerically, which can be further processed by a classification algorithm. This step provides a crucial bridge between images as visual data and machine learning algorithms that operate on numbers.

2.3 Gray Level Co-Occurrence Matrix (GLCM)

GLCM is one of the most popular feature extraction methods in image analysis, especially for describing texture. According to previous research, GLCM can help describe changes or variations that occur on the surface of an object [14]. This technique works by creating a new matrix containing frequency patterns from the original image data [15]

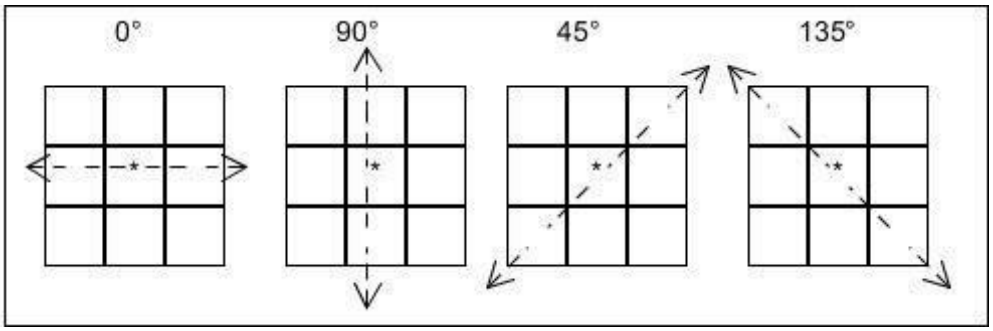


Figure 6. GLCM Angle

In GLCM, each pair of pixels has a specific angular orientation (θ) and distance (d) displayed in pixels or degrees [16]. In this study, GLCM parameters were set with pixel distance d = 1 and angles θ = [0°, 45°, 90°, 135°], gray levels = 8. For each image, four statistical attributes—Contrast, Correlation, Energy, and Homogeneity—were extracted at each orientation and averaged to form the final feature vector [17]. The selection of these four angles aims to make the texture representation richer and not limited to one particular direction. Thus, the classification model is expected to be able to recognize weather patterns from various visual perspectives.

2.4 Splitting Data

The next stage is data splitting. The processed data is then randomly divided into two parts: the training set and the testing set. This method allows the model to be trained before testing its performance on new data. This concept has long been used in machine learning because it can reduce bias in the model [18].

This data separation usually uses a certain ratio, such as 90:10, 80:20, or 70:30, depending on the research needs [19]. The choice of ratio greatly affects the model results because if the

training data is too little, the model may not be able to recognize patterns well. Conversely, if the test data is too little, the model evaluation results will be less representative [20].

In addition to the usual data separation, this study also used the Cross Validation (CV) technique to ensure the model does not experience overfitting. CV is a model evaluation method used to estimate performance on new, previously unseen data [21]. CV is considered important because it can reduce bias while ensuring the model is not only good on training data, but also reliable when faced with real data [22]. If a model has a low error rate on training data but a high one on test data, then it is certain that the model is experiencing overfitting [23].

2.5 Classification using Random Forest (RF) Algorithm

The Random Forest (RF) algorithm is an ensemble method consisting of many decision trees. Each decision tree in RF works by classifying data into a specific class. The final result is obtained by combining predictions from all trees through a majority voting method, thus significantly increasing accuracy [24].

The process of forming a decision tree in RF is almost the same as the Classification and Regression Trees (CART) method, only RF does not use the pruning process [25]. The main steps in building RF include: (a) taking bootstrap samples from the dataset, (b) building a decision tree from each sample by randomly selecting variables at each node, (c) voting for new data predictions, and (d) calculating the error rate using data not included in the bootstrap (out-of-bag error). With this combination of steps, RF becomes one of the reliable algorithms in handling image data with high variation.

2.6 Classification using the CatBoost Algorithm

In addition to RF, this study also uses the CatBoost algorithm. CatBoost is an implementation of Gradient Boosting that uses decision trees as a basis for prediction [26]. The main advantage of CatBoost lies in its ability to handle categorical data directly without the need for additional encoding processes. This makes CatBoost more efficient than other boosting algorithms [27].

CatBoost also adopts two important innovations: Ordered Target Statistics and Ordered Boosting, which reduce bias and overfitting. With this approach, CatBoost not only corrects errors from previous models but also maintains model stability when faced with new data. Another advantage of CatBoost is its internal regularization technique, which effectively prevents models from becoming overly complex. Therefore, CatBoost is often chosen for classification problems that require high accuracy with varied data.

2.7 Evaluation

The final stage in this research is model evaluation. After the model is built using RF and CatBoost, its performance will be tested using accuracy metrics and visualized using a Confusion Matrix. The Confusion Matrix is a very popular evaluation tool because it can provide detailed information about the model's ability to classify each class [28].

Unlike a single metric like accuracy, the Confusion Matrix shows the model's performance on each class separately. For example, how well the model recognizes rainy images compared to sunny images. This allows researchers to identify specific weaknesses in the model being built [29]. This type of evaluation is crucial to ensure that the model not only has high overall accuracy but also fairness in predicting all existing weather classes [30].

False Positive (FP), True Negative (TN), True Positive (TP), and False Negative (FN) are the four main cells in the confusion matrix. False Positive (FP) is the number of samples incorrectly classified as positive, False Negative (FN) is the number of samples incorrectly classified as negative, True Positive (TP) represents the number of samples correctly classified as positive, and True Negative (TN) is the number of samples correctly classified as negative [31]. Accuracy, precision, recall, and F1 score are some of the formulas related to the confusion matrix and are expressed as follows:

1. Accuracy

Accuracy is a measure of how accurately a model classifies all data samples. It is calculated as the ratio of the number of correct predictions (TP and TN) to the total number of samples.[32]. The following equation can be used to calculate accuracy.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FN + FP + TN}} \quad \text{..... (1)}$$

2. Precision

Precision is a measure of how many of all positive predictions are actually correct. It is calculated as the ratio between the number of correct positive predictions and the number of incorrect positive predictions [33]. Precision can be calculated using the following equation.

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

3. Recall

Recall, also called sensitivity, is a measure of how many of the total positive classes a model successfully predicts. Recall is calculated as the ratio of true positives to the total number of positive classes, which consists of true positives and false negativesv[34]. Recall calculations can be done using the following equation.

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

4. F1-Score

The F1-Score is calculated as the harmonic mean of precision and recall, giving the overall performance of a classification model by taking into account the balance between the two. It is a composite metric that combines precision and recall [35]. The F1 value can be calculated using the following equation.

$$\boldsymbol{F1 - Score} = \boldsymbol{2} * \frac{\boldsymbol{Precision * Recall}}{\boldsymbol{Precision + Recall}} \quad \text{..... (4)}$$

3. RESULTS AND DISCUSSION

The discussion can be made in several sub-chapters. In the Research on the Comparison of Weather Classification Methods on Weather Images with Gray Level Co-Occurrence Matrix (GLCM) Feature Extraction using Random Forest and CatBoost Algorithms aims to gain an in-depth understanding of the effectiveness of two different classification algorithms in processing weather image data. This topic is important because the current weather information system is not only used for academic purposes, but also has a major impact on people's daily activities, such as transportation, agriculture, fisheries, and disaster mitigation. With the presence of machine learning-based classification methods, it is hoped that the weather analysis process can be carried out faster and more accurately, thereby helping in strategic decision making.

In the initial stage of the research, weather image pre-processing was carried out using the GLCM (Gray Level Co-Occurrence Matrix) method. This method was chosen because it is able to extract image texture features well through main attributes such as Contrast, Correlation, Energy, and Homogeneity [22]. Contrast measures the difference in intensity between adjacent pixels, Correlation measures the linear relationship between pixels, Energy indicates the regularity or repetition of patterns in the image, while Homogeneity indicates the level of uniformity of intensity distribution. By calculating these four attributes at various angles (0°, 45°, 90°, 135°), more detailed information about image texture patterns can be obtained. The Gray Level Co-Occurrence Matrix (GLCM) was used to extract texture features based on spatial relationships between pixels. Parameters were empirically determined through preliminary experiments and literature review Distance (d) 1 pixel (balances local texture sensitivity and computation), Angles (θ) 0°, 45°, 90°, 135° to capture directional texture variance, Gray levels 8 (quantization reduces noise and

computation). Four GLCM statistical attributes were computed for each orientation Contrast, Correlation, Energy, and Homogeneity and their mean values were used as final features.

The results of the feature extraction are then categorized according to the type of weather based on the range of values formed. For example, Label 1 is used for the type of weather Cloudy, Label 2 for Rain, Label 3 for Shine, and Label 4 for Sunrise. This grouping process is very important so that the image has a clear class identity before being used in the classification process.

Table 1. GLCM Feature Extraction Results

Data	Angle	GLCM Extraction Features			
		Contrast	Correlation	Energy	Homogeneity
Cloudy.jpg	0°	0.03238313856	0.9944295698	0.1621684122	0.9838139955
	45°	0.05164413538	0.991107271	0.1569785376	0.9741890899
	90°	0.03808688388	0.9934357029	0.160588748	0.9809732665
	135°	0.04899142681	0.9915640108	0.1578642158	0.9755879683
Rain.jpg	0°	0.3065923863	0.8285951289	0.1642821407	0.8526823481
	45°	0.2139669259	0.8802651608	0.1904421442	0.8933064792
	90°	0.1737505163	0.902719181	0.2054286466	0.9133174997
	135°	0.4023671703	0.7748373107	0.1452201354	0.8151817488
Shine.jpg	0°	0.06112365623	0.9902963283	0.1497532038	0.9717999424
	45°	0.1355886528	0.978464016	0.1355991423	0.9438058644
	90°	0.1183035025	0.9812091377	0.1376618473	0.948405674
	135°	0.1353143363	0.9785075751	0.1359150605	0.9440970279
Sunrise.jpg	0°	0.03337249666	0.9944875285	0.1816186586	0.9847680018
	90°	0.07270883556	0.9879844841	0.1721160659	0.9700520311
	90°	0.06629525718	0.9890395972	0.1732803744	0.9722444699
	135°	0.0783917635	0.9870453505	0.1714068829	0.9688234056

After feature grouping, data splitting is performed using the cross-validation method (5-Fold CV). In this study, an 80:20 ratio was used, with 80% as training data and 20% as test data. This ratio is considered quite ideal because it provides a large enough training data to train the model, but still leaves sufficient test data to evaluate model performance. In addition, cross-validation plays an important role in avoiding overfitting, which is a condition where the model is too good at recognizing training data but fails when faced with new data [28]. With this technique, the reliability of the model can be tested objectively.

Table 2. Distribution of Training Data and Test Data

Data	Ratio (%)
Training Data	80%
Testing Data	20%
Total	100%

After data division, a classification process is carried out using the Random Forest and CatBoost algorithms. Random Forest works by forming a large number of independent decision trees, then the classification results are determined based on majority voting. The advantages of this method are its stability in dealing with large and varied datasets, as well as its ability to reduce classification errors through an ensemble approach.

In contrast, CatBoost is a variant of the gradient boosting algorithm designed to address the weaknesses of other boosting methods, particularly in handling categorical data. CatBoost uses Ordered Target Statistics and Ordered Boosting techniques to address the weaknesses of traditional gradient boosting. However, in the case of texture-based weather imagery, CatBoost's performance may not be as optimal as Random Forest due to the more complex nature of the data.

Random Forest was trained with hyperparameters: n_estimators = 200, max_depth = 12, and random_state = 42. CatBoost was trained with iterations = 100, learning_rate = 0.1, and depth = 6. Both models were evaluated using 5-fold cross-validation. Training was conducted on 80% of the dataset, with 20% reserved for testing.

Based on the test results, the Random Forest algorithm with the parameter random_state = 42 was able to achieve an accuracy of 92.43% (± 0.0116). These results indicate that Random Forest has excellent performance in classifying weather images.

Table 3. Random Forest Classification Results (5-Fold CV)

Random Forest (random_state)	Accuracy
42	92.43% (± 0.0116)

Meanwhile, the CatBoost algorithm with iterations = 100 parameter produces an accuracy of 68.88%, lower than Random Forest [34]. Nevertheless, CatBoost still has the advantage of more efficient handling of categorical data.

Table 4. Results of Catboost Classification (5-Fold CV)

Catboost (iterations)	Accuracy
100	68.88% (± 0.0116)

To further evaluate performance, the Confusion Matrix is used as a measuring tool. The Confusion Matrix not only shows the level of accuracy, but also shows precision, recall, and F1-score which describe the model's ability to classify each class in more detail.

The evaluation results show that Random Forest has an accuracy value of 0.9243 ± 0.0116, precision of 0.9250 ± 0.0117, recall of 0.9243 ± 0.0116, and F1-score of 0.9242 ± 0.0117. These figures prove that Random Forest works consistently and balanced in recognizing the four types of weather.

Table 5. Confusion Matrix Random Forest Evaluation Results (5-Fold CV)

Accuracy	Precision	Recall	F1-Score
0.9243 ± 0.0116	0.9250 ± 0.0117	0.9243 ± 0.0116	0.9242 ± 0.0117

On the other hand, CatBoost produced accuracy values of 0.6888 ± 0.0176, precision of 0.6866 ± 0.0177, recall of 0.6888 ± 0.0176, and F1-score of 0.6848 ± 0.0160. These results indicate a significant performance gap compared to Random Forest.

Table 6. CatBoost Confusion Matrix Evaluation Results (5-Fold CV)

Accuracy	Precision	Recall	F1-Score
0.6888 ± 0.0176	0.6866 ± 0.0177	0.6888 ± 0.0176	0.6848 ± 0.0160

In addition to tables, the research results are also visualized using a Confusion Matrix Plot, which shows the distribution of classification errors in each class. Figure 9 shows the plot for Random Forest, while Figure 10 shows the plot for CatBoost.

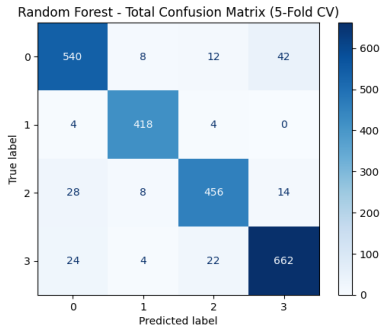


Figure 9. Confusion Matrix Random Forest (5-Fold CV)

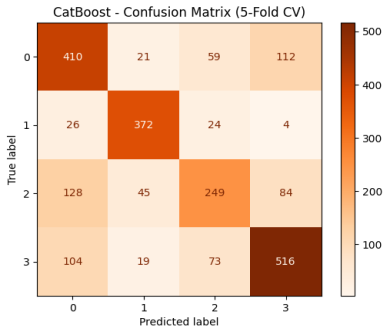


Figure 10. CatBoost Confusion Matrix (5-Fold CV)

From this visualization, it can be seen that Random Forest tends to make fewer classification errors, especially for the Cloudy and Rain classes, which are visually similar. This

demonstrates that bagging-based ensemble methods like Random Forest are capable of addressing the challenges of image data with high texture variation.

Overall, the results of this discussion indicate that the Random Forest algorithm is superior to CatBoost in weather image classification using GLCM feature extraction. However, CatBoost still has potential for further development, especially with parameter adjustments and the use of more complex optimization techniques. This study emphasizes that the choice of machine learning algorithm must be tailored to the characteristics of the data being processed. For texture-based data such as weather imagery, ensemble decision tree-based methods such as Random Forest have proven more effective. showing higher variance and less stability. The 5-fold CV confirmed that RF consistently outperformed CatBoost across all folds with minimal fluctuation (<2%), demonstrating stronger generalization.

4. CONCLUSION

Based on the results and discussion on Comparative Weather Classification Methods on Weather Images Using Gray Level Co-Occurrence Matrix (GLCM) Feature Extraction with Random Forest and CatBoost Algorithms, it can be concluded that the use of the GLCM feature extraction method in processing weather images provides an important contribution in producing texture attributes that are able to represent the characteristics of an image. The four main attributes produced, namely Contrast, Correlation, Energy, and Homogeneity, have proven effective in describing differences in weather image patterns, such as cloudy, rainy, sunny, or sunrise. With this pre-processing process, the classification model has a more structured and representative data foundation for further testing.

Furthermore, the data division process using the cross-validation (5-Fold CV) method with a ratio of 80% training data and 20% test data is a crucial step in maintaining the validity of the classification results. Through algorithm comparisons, it was found that the Random Forest algorithm provided far more optimal results than CatBoost. Random Forest with the random_state parameter = 42 managed to achieve an accuracy level of 92.43% (± 0.0116), with relatively balanced precision, recall, and F1-score values of around 92%. This indicates that the bagging method used by Random Forest is able to overcome data variability and produce a stable model with a low error rate.

In contrast, the CatBoost algorithm with iterations = 100 only achieved an accuracy of 68.88% (± 0.0176), with precision, recall, and F1-score values around 68%. This difference indicates that although CatBoost is known as one of the superior boosting algorithms in handling categorical variables, in the context of weather imagery based on GLCM texture features, its performance is not as strong as Random Forest.

Thus, it can be concluded that the combination of the GLCM feature extraction method and the Random Forest algorithm is the most effective approach for classifying weather imagery. The results of this study also emphasize the importance of selecting the appropriate classification method according to the characteristics of the data used. Future research can be expanded by adding parameter variations, using larger datasets, or combining other algorithms to obtain more accurate and applicable weather classification results in modern weather information systems.

REFERENCES

[1] T. Lattifia, P. Wira Buana, N. Kadek, and D. Rusjayanthi, “Model Prediksi Cuaca Menggunakan Metode LSTM,” 2022.

[2] M. A. Siregar, Tukino, S. Faisal, A. Fauzi, and I. Kadori, “Klasifikasi untuk Prediksi Cuaca Menggunakan Asemble Learning,” 2020.

[3] F. Amir, E. Utami, and H. Hanafi, “Literature Study on the Development of Neural Networks For Weather Forecasting,” *J Teknol*, vol. 17, no. 1, pp. 49–57, Jun. 2024, doi: 10.34151/jurtek.v17i1.4637.

[4] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, “Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 187–192, Feb. 2021, doi: 10.29207/resti.v5i1.2813.

[5] S. Barburiceanu, R. Terebes, and S. Meza, "3d texture feature extraction and classification using glcm and lbp-based descriptors," *Applied Sciences (Switzerland)*, vol. 11, no. 5, pp. 1–26, Mar. 2021, doi: 10.3390/app11052332.

[6] A. Suresh, R. Udendhran, and M. Balamurgan, "Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers," *Soft comput*, vol. 24, no. 11, pp. 7947–7953, Jun. 2020, doi: 10.1007/s00500-019-04066-4.

[7] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[8] Q. Meidianingsih and D. Agustine, "Study of Bagging Application in the Safe-Level Smote Method in Handling Unbalanced Classification," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 105–116, Mar. 2021, doi: 10.29244/ijsa.v5i1p105-116.

[9] S. C. Hsu, A. K. Sharma, R. Tanone, and Y. T. Ye, "Predicting Rainfall Using Random Forest and CatBoost Models," in *World Congress on Civil, Structural, and Environmental Engineering*, Avestia Publishing, 2024. doi: 10.11159/icgre24.146.

[10] G. Ngo, R. Beard, and R. Chandra, "Evolutionary bagging for ensemble learning," Aug. 2022, doi: 10.1016/j.neucom.2022.08.055.

[11] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," 2020, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/JSTARS.2020.3026724.

[12] Y. Miftahuddin, S. UMAROH, and A. M. YAMANI, "Peningkatan Random Forest dengan menerapkan GLCM (Gray Level Co-Occurence Matrix) pada Klasifikasi Leaf Blast Tumbuhan Padi," *MIND Journal*, vol. 7, no. 1, pp. 37–50, Jun. 2022, doi: 10.26760/mindjournal.v7i1.37-50.

[13] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P.-H. Huynh, "Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 125–136, Mar. 2024, doi: 10.35882/jeeemi.v6i2.382.

[14] N. Iqbal, R. Mumtaz, U. Shafi, and S. M. H. Zaidi, "Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms," *PeerJ Comput Sci*, vol. 7, p. e536, May 2021, doi: 10.7717/peerj-cs.536.

[15] M. Teke and T. Etem, "Cascading GLCM and T-SNE for detecting tumor on kidney CT images with lightweight machine learning design," *Eur Phys J Spec Top*, vol. 234, no. 15, pp. 4619–4634, Oct. 2025, doi: 10.1140/epjs/s11734-025-01842-2.

[16] K. Djunaidi, H. Bedi Agtriadi, D. Kuswardani, and Y. S. Purwanto, "Gray level co-occurrence matrix feature extraction and histogram in breast cancer classification with ultrasonographic imagery," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 795, May 2021, doi: 10.11591/ijeecs.v22.i2.pp795-800.

[17] I. Sutrisno *et al.*, "Design of Pothole Detector Using Gray Level Co-occurrence Matrix (GLCM) And Neural Network (NN)," *IOP Conf Ser Mater Sci Eng*, vol. 874, no. 1, p. 012012, Jun. 2020, doi: 10.1088/1757-899X/874/1/012012.

[18] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[19] J.-H. Syu, M. Fojcik, R. Cupek, and J. C.-W. Lin, "HTTPS: Heterogeneous Transfer learning for spliT Prediction System evaluated on healthcare data," *Information Fusion*, vol. 113, p. 102617, Jan. 2025, doi: 10.1016/j.inffus.2024.102617.

[20] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Math Probl Eng*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.

[21] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3182.

[22] G. Valente, A. L. Castellanos, L. Hausfeld, F. De Martino, and E. Formisano, "Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes," *Neuroimage*, vol. 238, p. 118145, Sep. 2021, doi: 10.1016/j.neuroimage.2021.118145.

[23] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Comput Stat*, vol. 36, no. 3, pp. 2009–2031, Sep. 2021, doi: 10.1007/s00180-020-00999-9.

[24] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.

[25] S. Jayaprakash, S. Krishnan, and J. V, *Predicting Students Academic Performance using an Improved Random Forest Classifier*. IEEE, 2020.

[26] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 2, Aug. 2022, doi: 10.28932/jutisi.v8i2.4906.

[27] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[28] K. Xie, L. Huang, W. Zhang, Q. Qin, and L. Lyu, "A CNN-based multi-task framework for weather recognition with multi-scale weather cues," *Expert Syst Appl*, vol. 198, p. 116689, Jul. 2022, doi: 10.1016/j.eswa.2022.116689.

[29] C.-L. Fan, "Evaluation Model for Crack Detection with Deep Learning: Improved Confusion Matrix Based on Linear Features," *J Constr Eng Manag*, vol. 151, no. 3, Mar. 2025, doi: 10.1061/JCEMD4.COENG-14976.

[30] W. Ren, X. Xiao, J. Xu, H. Chen, Y. Zhang, and J. Zhang, "Trojan Virus Detection and Classification Based on Graph Convolutional Neural Network Algorithm," *Journal of Industrial Engineering and Applied Science*, vol. 3, no. 2, pp. 1–5, Apr. 2025, doi: 10.70393/6a69656173.323735.

[31] Z. Zhang *et al.*, "Improving text classification via computing category correlation matrix from text graph," *Comput Speech Lang*, vol. 89, p. 101688, Jan. 2025, doi: 10.1016/j.csl.2024.101688.

[32] R. A. A. Malik and Y. Sibaroni, "Multi-aspect Sentiment Analysis of Tiktok Application Usage Using FasText Feature Expansion and CNN Method," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 277–285, Sep. 2022, doi: 10.47065/josyc.v3i4.2033.

[33] M. A. Rafif, G. Sanjaya Indrajaya, M. K. Al-Ghazi, J. Johnny, and N. T. M. Sagala, "Comparison of Decision Tree and Support Vector Machine for Predicting Jakarta Air Quality Index," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, Feb. 2023, pp. 381–385. doi: 10.1109/ICCoSITE57641.2023.10127855.

[34] N. Eliyati, M. Rahmayani, S. Wijaya, D. A. Zayanti, E. S. Kresnawati, and Y. Resti, "Prediction Of Air Quality Index Using Decision Tree With Discretization," *Indonesian Journal of Engineering and Science*, vol. 3, no. 3, pp. 061–067, Nov. 2022, doi: 10.51630/ijes.v3i3.82.

[35] I. M. Putra, I. Tahyudin, H. A. Awal Rozaq, A. Yahya Syafa'at, R. Wahyudi, and E. Winarto, "Classification Analysis of COVID19 Patient Data at Government Hospital of Banyumas using Machine Learning," in *2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, IEEE, Jun. 2021, pp. 271–274. doi: 10.1109/ICSCEE50312.2021.9498020.