Analysis of Synthetic Data Utilization with Generative Adversarial Network in Flood Classification using K-Nearest Neighbor Algorithm

Wahyu Afriza, Mardhani Riasetiawan*, Dyah Aruming Tyas
Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia

Abstract*—Indonesia is a country with a tropical climate that has high rainfall rates and is supported by the uncertainty of weather and climate conditions. With the uncertainty of weather and climate as well as flood events, minimal predictive information on flooding, and the lack of availability of data on the causes of flooding, a comparison of synthetic data generation from the minimal data available from BMKG with synthetic data generation from Kaggle online platform data in the form of temperature and humidity data, rainfall, and wind speed from BMKG and annual rain data from Kaggle was analyzed. This research aims to obtain the results of data comparison analysis of synthetic data generation from different datasets with the benchmark of classification system results using K-Nearest Neighbor (KNN) and accuracy evaluation with Confusion Matrix. The research process uses climate data from the BMKG DI Yogyakarta Climatology Station within 20 months, the Geophysical Station within 12 months, and Kerala data with a range of 1901–2018. Synthetic data generation is done using the Conditional Tabular Generative Adversarial Network (CTGAN) model. CTGAN produces quite good data in terms of distribution and data differences if the original data is large and the synthetic data generated is small. The KNN classification system on the BMKG data experienced overfitting, as indicated by the accuracy value in the evaluation increasing in the range of 85–94% and the validation decreasing in the range of 89%–65%. This is because there is no uniqueness in the data and too little original data made into synthetics, which affects the difficulty of the classification system in identifying data that is quite different in distance and data values generated by CTGAN. In Kerala, the accuracy value on evaluation is in the range of 92–95%, and validation is in the range of 0.7–0.83%, with Classifier k1 being the most optimal system.

***Keywords*—Classification; rainfall; synthetic data; KNN; GAN**

I. Introduction

Indonesia is a tropical country located in the equatorial region with high rainfall. Today's climate change is affecting the weather and climate to the extent that high water discharge causes flooding. Water discharge can be caused by heavy rains with short or long durations, as in the rains that hit the Asian sub-continent with the deadliest floods that damage the environment, agricultural land, and basic health facilities [1]. A flood is a state of water inundation for a certain time, even though it is in an area that rarely floods with the support of rainfall and a long duration [2]. Floods can affect living things, wind pressure, temperature, watercolor, wind direction, humidity, and more, and have the potential to damage property and buildings [3], as well as adversely affect human health, the environment, cultural heritage, and economic activities [4].

The characteristics of heavy rainfall are strongly influenced by spatiotemporal patterns, space- and time-based models, and the amount of rainfall. Location: with damage intensity within the watershed, damage patterns (flooding from rivers, flooding from inland waters, sediment-related disasters, and other) vary depending on the distribution of rainfall. In order to implement effective flood control measures, it is important to understand the rainfall patterns that occur in the watershed and take countermeasures based on the characteristics of the associated hazards [5].

Some of the factors that affect the occurrence of floods are temperature, humidity, dew point temperature, wind speed, river flow volume, water level, and rainfall volume. The amount of rainfall is a major factor in the hydrological cycle process by monitoring the balance of freshwater and saltwater resources. Process of data acquisition can be done with the use of the Internet of Things based on data from the sensors used. Rainfall prediction or forecasting plays an important role in hydrological modeling and management of water resource issues such as flood warnings and real-time control of urban drainage systems [6].

In this research, a comparative analysis of the use of synthetic data in making a classification system based on the machine learning algorithm K-Nearest Neighbor (KNN) is carried out. Synthetic data generation is carried out due to the lack of availability and types of data features that can be obtained from BMKG Online Data (https://dataonline.BMKG.go.id/) and open data from online platforms for the classification of flood disaster events. The data used is BMKG data with rainfall data parameters, temperature and humidity, wind speed, and flood events as benchmarks for measuring and determining potential flood classes as well as monthly and annual flood data. This research is intended as an analysis of the use of synthetic data on climate data and natural disasters.

II. Research Methodology

A. System Needs Analysis

There are several stages in designing a classification system, including design, data preparation, training, and testing of a classification system based on the K-Nearest Neighbor (KNN) algorithm. The data used is BMKG data as training and

validation data and Kerala data. Dataset creation includes downloading data and merging BMKG and BPBD data to become BMKG data with data that has a flood class label. The flood level data entry is in accordance with Table I. Then, the Kerala data underwent a download process without any additional processing.

TABLE I. FLOODING LEVEL IN YOGYAKARTA

No.	Flood Level	Flood High
1	Tidak/No	0 cm
2	Ringan/Low	< 100 cm
3	Tinggi/High	≥ 100 cm

BMKG training data has a time span of twenty months starting from January 2022 to August 2023; BMKG validation data has a time span of twelve months or one year starting from October 2022 to September 2023; and Kerala data in the form of rainfall data and annual flood classes has a time span of 1901–2018.

The BMKG training data will be divided into a 3:1 ratio for training and testing, so that 75% of the data will be used in the training process and 25% of the data will be used in the testing process, which can be used as confusion matrix-based evaluation results. The BMKG data has a total of 320 rainfall data points over a time span of twenty months. In the validation of BMKG data, the data is fully used as validation of the classification system results. Furthermore, Kerala data totals 118 data points, which will be divided into 100 training data points with the same division as BMKG data, namely 75% and 25%, and 18 data points as validation data from the Kerala classification system.

B. Synthetic Data Generation

Synthetic data is artificial data generated from the original data. Synthetic data can overcome the problems of data security, data confidentiality, unbalanced data, and others. The generative adversarial network works based on two neural networks: the discriminator and the generator [7]. GAN has several mathematical formulas for calculations. In GAN, there is a discriminator in Eq. (1), a generator in Eq. (2), and training for the discriminator and generator is shown in Eq. (3) and Eq. (4) [8].

$$L_D = Error(D(x), 1) + Error\big(D\big(G(z)\big), 0\big) \quad (1)$$

$$L_G = Error\big(D\big(G(z)\big), 1\big) \quad (2)$$

$$V(G,D) = E_{x\sim P_{data}}\big[\log\big(D(x)\big)\big] + E_{z\sim P_z}[\log(1 - D(G(z)))] \quad (3)$$

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \quad (4)$$

$$V(G,D^*) = E_{x\sim P_{data}}\big[\log\big(D^*(x)\big)\big] + E_{x\sim P_g}\big[\log\big(1 - D^*(x)\big)\big] \quad (5)$$

C. Classification Method Implementation

After obtaining data from the source, the data will undergo a merging process between climate data and flood event data, then data cleaning will be carried out from unnecessary data or data with empty values and 8888 (unmeasured data) by manipulating the data using the median value.

The KNN method will use BMKG data which will be divided into train and test data. After defining the data, we will look for the minimum and maximum distance from the calculation of the train data distance and then the minimum distance from the calculation of the test data distance to the train data which will be assigned to several flood classes. The data will be classified into three flood classes, namely the No, Mild, and High classes as shown in Fig. 1.

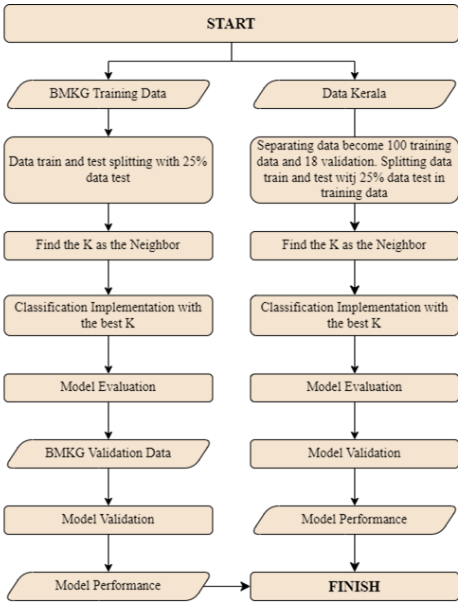

Fig. 1. Classification implementation.

D. Evaluation and Validation

The system evaluation process will be carried out by testing whether the system is able to classify rainfall, temperature, humidity, and wind speed data that can potentially flood into three flood classes. The measurement will be carried out by utilizing the Confusion Matrix Theory, which will compare the output of the system with the actual label of the data and will then produce accuracy, precision, Recall, and f1-score numbers according to Eq. (6) to Eq. (11) In the validation process, the same thing is done but with different data, namely data that has never experienced the train and test process.

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+TN} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$FI - Score = \frac{1}{\frac{1}{recall} + \frac{1}{presisi}} \quad (9)$$

III. RESULT AND DISCUSSION

The datasets used in the research are from BMKG data and Kerala data, which are then divided into train, test, and validation data. BMKG data are gather from the filed sensor in the real situation. Kerala data will later be used in making a classification system with a validation process from different data and the training process that has been carried out. BMKG data that has been downloaded from the BMKG and BPBD

DIY online portals has a distribution of rain event data totaling 320 and 159 in BMKG training and validation data, respectively. The cleaning process is done by deleting the non-rain values in the BMKG data. In Kerala data, no cleaning was done because all data will be used.

After the non-rain data was eliminated, data manipulation was performed to fill in the values of the variable components in each column that were zero, null, and unmeasured except RR. The review for data manipulation was conducted only on the Tavg, RH_avg, and ff_avg features. Referring to Fig. 2, the data distribution on each feature except rainfall data (RR) is unevenly skewed with the category "skewed negative," or the mean value is lower than the median value of the data. In this situation, data manipulation using the median value aims to direct the distribution to a normal distribution. This was also done on the validation BMKG data.

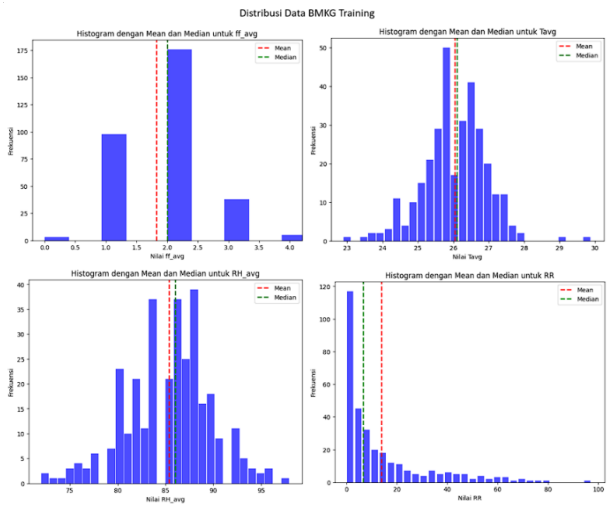

Fig. 2. Distribution of feature data.

After doing data manipulation on data rows that are 0, null, and unmeasured, as shown in Table II, In Kerala data, there are two flood classes, namely YES and NO as shown in Table III.

In the BMKG training data that has been processed, the data is not balanced between classes. If the process of making a classification system using this data is not followed, the result of the classification system will produce a poor system and undertraining, a situation when the classification system can recognize one class well but cannot recognize the other class, which tends to result in the classification system that has been made classifying data for the majority class [9]. This can be overcome by multiplying existing data with generative data or synthetic data.

The data generation process is carried out with CTGAN. This was done to prevent the potential for an undertrained classification system. The generative process was carried out with five experiments with different amounts of data. The flood class in the data will be converted into an integer or number, with 0 as no flood, 1 as a minor flood, and 2 as a high flood. The original flood data, with a total of six minor flood events and one high flood event, doubled without changes to two for the model calculation process, will be trained by GAN and generate synthetic data.

TABLE II. TOTAL BMKG DATA IN EACH CLASS

No.	Data	Flood	Case
1	BMKG Training	No	313
2		Mild	6
3		High	1
4	BMKG Validation	No	156
5		Mild	2
6		High	1

TABLE III. TOTAL KERALA DATA IN EACH CLASS

No.	Data	Flood	Case
1	Kerala Training	YES	52
2		NO	48
3	Kerala Validation	YES	8
4		NO	10

Furthermore, synthetic data was created for each data point. In BMKG data, synthetic data for mild flood and high flood classes is made into 30, 60, 90, 120, and 150 data points. In Kerala, synthetic data for yes and no classes was made into 150 of all the data and only 6 data samples. The distribution and differences between synthetic and real data can be seen in Fig. 3 to Fig. 14.

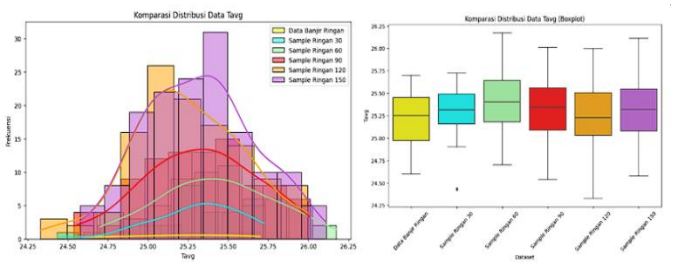

Fig. 3. Distribution of Tavg in mild class.

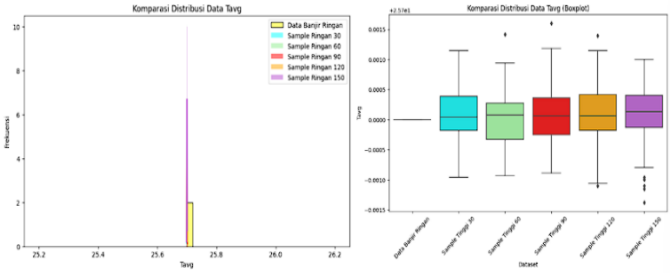

Fig. 4. Distribution of Tavg in high class.

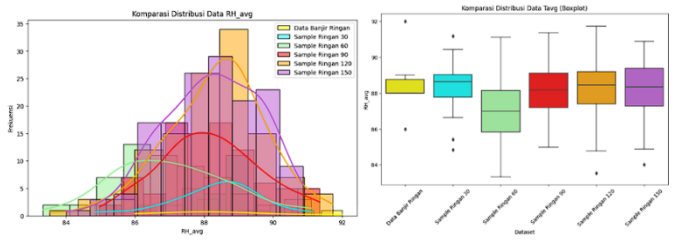

Fig. 5. Distribution of RH_avg in mild class.

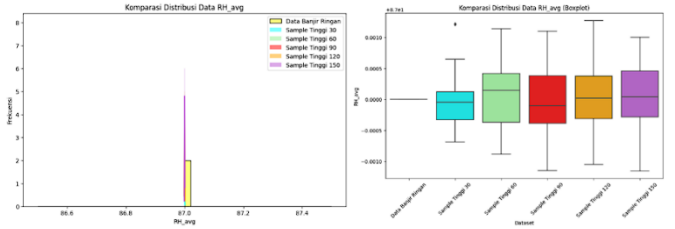

Fig. 6. Distribution of RH_avg in high class.

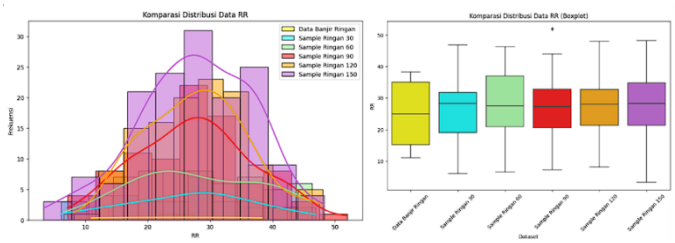

Fig. 7. Distribution of RR in mild class.

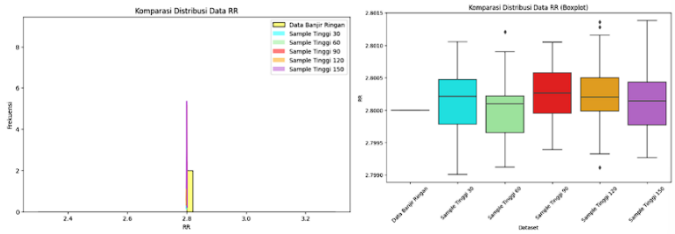

Fig. 8. Distribution of RR in high class.

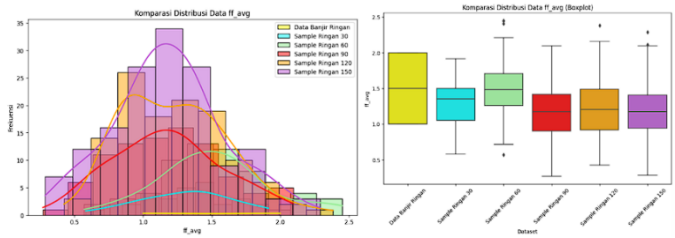

Fig. 9. Distribution of ff_avg in mild class.

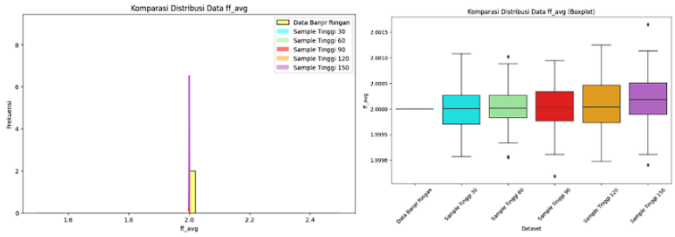

Fig. 10. Distribution of ff_avg in high class.

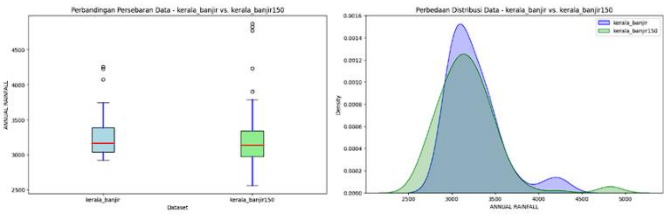

Fig. 11. Distribution of flood data in Kerala real data to 150

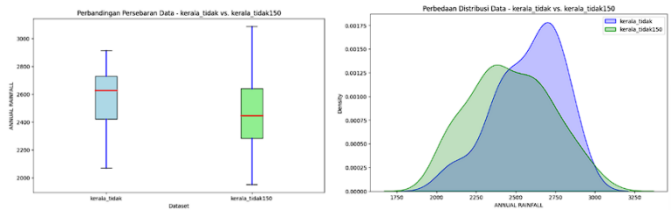

Fig. 12. Distribution of not flood data in Kerala real data to 150

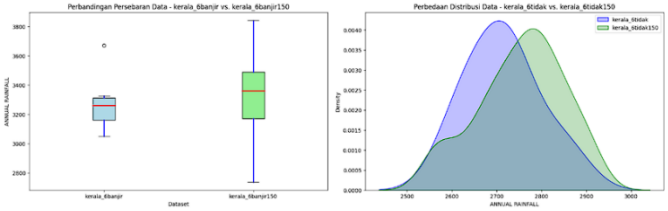

Fig. 13. Distribution of flood data in Kerala 6 real data to 150

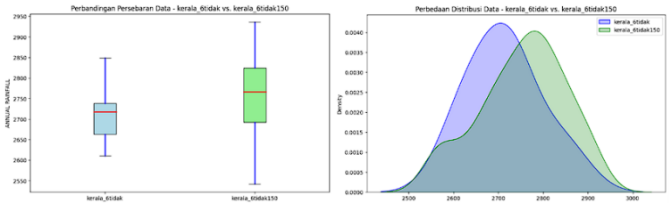

Fig. 14. Distribution of not flood data in Kerala 6 real data to 150

TABLE IV. TOTAL SYNTHETIC DATA OF BMKG DATA

No.	Model	No Class	Mild Class	High Class
1	*Classifier* 1	313	36	32
2	*Classifier* 2	313	66	62
3	*Classifier* 3	313	96	92
4	*Classifier* 4	313	126	122
5	*Classifier* 5	313	156	152

TABLE V. TOTAL SYNTHETIC DATA OF KERALA DATA

No.	Model	YES	NO
1	*Classifier k*	52	48
2	*Classifier k1*	202	198
3	*Classifier k2*	156	156

Before making a classification system in each experiment based on the data that shown in Table IV and Table V, the K value is determined as a consideration for determining neighbors in calcification using the Euclidean distance, which has a calculation formula as in Eq. (10) and Eq. (11).

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (10)$$

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (11)$$

In the process of determining the value of K, the best value was found using repetition operations throughout the classifier experiment. Based on the accuracy results, a K value of 5 is obtained with the consideration of a good and consistent accuracy value, with an average accuracy value of 91.38% on BMKG data and 92.26% on Kerala data.

The classification system that has been created in each experiment will be evaluated to determine and assess the ability or performance of the system. In the system evaluation, each classifier is evaluated by utilizing the confusion matrix theory based on the formulas in Eq. (6) to Eq. (9). In the evaluation process, test data is used, which amounts to 25% of the total of all classes. The confusion matrix results of each classifier can be seen in Table VI for the BMKG data and Table VII for the Kerala data, as well as the comparison graph in Fig. 15.

TABLE VI. BMKG DATA CLASSIFICATION EVALUAION

No.	Model	Precission	Recall	F-1 score	Accuracy
1	*Classifier 1*	0.82	0.84	0.83	0.84
2	*Classifier 2*	0.91	0.91	0.91	0.91
3	*Classifier 3*	0.93	0.92	0.92	0.92
4	*Classifier 4*	0.93	0.92	0.93	0.92
5	*Classifier 5*	0.93	0.93	0.93	0.93

TABLE VII. KERALA DATA CLASSIFICATION EVALUATION

No.	Model	Precission	Recall	F-1 score	Accuracy
1	*Classifier k*	0.93	0.92	0.92	0.92
2	*Classifier k1*	0.91	0.90	0.90	0.90
3	*Classifier k2*	0.95	0.94	0.94	0.94

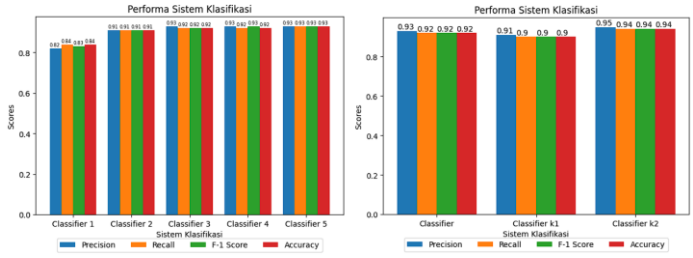

Fig. 15. Comparison of evaluation in each model.

Validation data on BMKG data has a total of 159 data points that have undergone data pre-processing. In the BMKG validation data, there are 156 non-flooding events, 2 minor floods, and 1 high flood. While in Kerala, the train data amounted to 18 data points, with 8 flood data points and 10 non-flood data points. In the validation test, the same process as the evaluation is carried out but with data that has never been trained and tested, utilizing the confusion matrix theory based on the formulas in Eq. (5), (6), (7), and (8) with the predicted data and the original label data. The results of the BMKG training data classification system validation can be seen in Table VIII, and the Kerala data system validation can be seen in Table IX. Then, the result comparison graph is in Fig. 16.

TABLE VIII. BMKG DATA CLASSIFICATION VALIDATION

No.	Model	Precission	Recall	F-1 score	Accuracy
1	*Classifier 1*	0.96	0.89	0.93	0.89
2	*Classifier 2*	0.96	0.86	0.91	0.86
3	*Classifier 3*	0.97	0.77	0.86	0.77
4	*Classifier 4*	0.97	0.74	0.83	0.74
5	*Classifier 5*	0.97	0.65	0.78	0.65

TABLE IX. KERALA DATA CLASSIFICATION VALIDATION

No.	Model	Precission	Recall	F-1 score	Accuracy
1	*Classifier k*	0.75	0.75	0.75	0.77
2	*Classifier k1*	0.78	0.88	0.82	0.83
3	*Classifier k2*	0.71	0.62	0.67	0.72

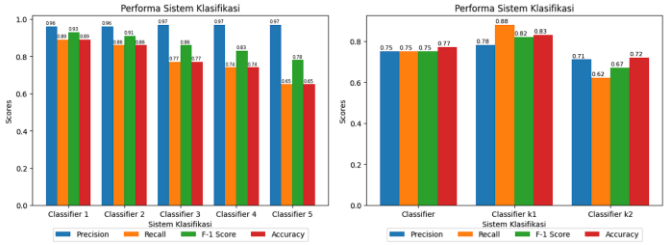

Fig. 16. Comparison of evaluation in each model

Referring to Fig. 17, the comparison of evaluation and validation results on BMKG data shows a significant difference in data, so the classification system from BMKG data has poor results, although Classifier 1 has the least difference. While in Kerala data, the evaluation and validation results are quite consistent in their improvement, with Classifier k1 being the most optimal, which is the creation of synthetic data from all the original Kerala data.

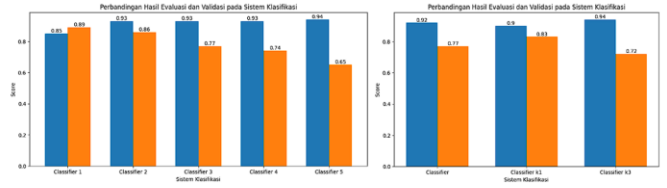

Fig. 17. Comparison of accuracy in evaluation and validation.

IV. CONCLUSION

The research indicates that the BMKG data, which includes temperature, humidity, rainfall, and wind speed features, does not have unique characteristics for each class. As a result, the classification system derived from this data suffers from overfitting, leading to imbalanced results between evaluation and validation. On the other hand, the Kerala data exhibits unique features for each class, which allows for a more accurate classification system. This system achieves an evaluation accuracy of 90-95% and a validation accuracy of 72-83%.

Synthetic data generated from Generative Adversarial Networks (GANs) can create a large amount of data from a small amount of original data. However, the quality of this

synthetic data is dependent on the quantity of original and synthetic data used. This can affect the similarity of the synthetic data to the original data, and vice versa. In terms of the classifiers, Classifier 1 (with 30 sample data), Classifier 2 (with 60 data samples), and Classifier k1 show similar accuracy values on evaluation and validation. However, Classifier 3, Classifier 4, Classifier 5, and Classifier k3 exhibit a significant difference in accuracy values on evaluation and validation.

In summary, the BMKG data's lack of unique class characteristics leads to overfitting in the classification system, resulting in imbalanced evaluation and validation results. In contrast, the Kerala data, with its unique class characteristics, produces a more accurate classification system. Synthetic data, generated from GANs, can be highly useful, but the quality of this data depends on the quantity of original and synthetic data used. Finally, the classifiers show varying accuracy values on evaluation and validation, with Classifiers 3, 4, 5, and k3 exhibiting significantly different results compared to Classifiers 1, 2, and k1. The analysis of synthetic data utilization with Generative Adversarial Network (GAN) in flood classification using the K-Nearest Neighbor (KNN) algorithm is an innovative and promising research area. To further advance this work and contribute to the field, the following future work suggestions are proposed Explore and develop more advanced GAN architectures to generate synthetic flood-related data. Investigate different GAN variants, such as Wasserstein GANs or Progressive GANs, to improve the quality and diversity of synthetic data. Conduct a thorough investigation into the hyperparameters of both the GAN and KNN algorithms to optimize their performance. This includes tuning learning rates, batch sizes, and other relevant parameters to achieve better results in terms of classification accuracy and computational efficiency. Extend the analysis by incorporating and comparing the performance of other machine learning models for flood classification. This could include algorithms like Support Vector Machines (SVM), Decision Trees, or Random Forests. A comprehensive comparison will provide insights into the strengths and weaknesses of different approaches. Test the proposed framework on a broader range of datasets to evaluate its generalizability. This includes datasets from different geographical locations, varied environmental conditions, and various types of flooding scenarios. Assess the robustness of the model across different contexts. By addressing these future work areas, the research can make significant contributions to the field of flood classification, synthetic data generation, and the intersection of GANs and KNN algorithms.

REFERENCES

[1] Babar, M., Rani, M. and Ali, I., 2022, November. A Deep learning-based rainfall prediction for flood management. In 2022 17th International Conference on Emerging Technologies (ICET) (pp. 196-199). IEEE.

[2] Al Kindhi, B., Triana, M.I., Yuhana, U.L., Damarnegara, S., Istiqomah, F. and Imaaduddiin, M.H., 2022, November. Flood Identification with Fuzzy Logic Based on Rainfall and Weather for Smart City Implementation. In 2022 IEEE International Conference on Communication, Networks, and Satellite (COMNETSAT) (pp. 67-72). IEEE.

[3] Khan, T.A., Shahid, Z., Alam, M., Su'ud, M.M. and Kadir, K., 2019, December. Early flood risk assessment using machine learning: A comparative study of svm, q-svm, k-nn, and lda. In 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS) (pp. 1-7). IEEE.

[4] Panganiban, E.B. and Cruz, J.C.D., 2017, November. Rainwater level information with flood warning system using flat clustering predictive technique. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 727-732). IEEE.

[5] Hoshino, T. and Yamada, T.J., 2023. Spatiotemporal classification of heavy rainfall patterns to characterize hydrographs in a high-resolution ensemble climate dataset. Journal of Hydrology, 617, p.128910.

[6] Adaryani, F.R., Mousavi, S.J. and Jafari, F., 2022. Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM, and CNN. Journal of Hydrology, 614, p.128463.

[7] Habibi, O., Chemmakha, M., & Lazaar, M. (2023). Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. Engineering Applications of Artificial Intelligence, 118, 105669.

[8] Kiran, A. and Kumar, S.S., 2023, March. A Comparative Analysis of GAN and VAE based Synthetic Data Generators for High Dimensional, Imbalanced Tabular data. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-6). IEEE.

[9] Karimi, Z., 2021. Confusion Matrix. Encycl. Mach. Learn. Data Min., no. October, pp.260-260.

[10] Freiesleben, T. and Grote, T., 2023. Beyond generalization: a theory of robustness in machine learning. Synthese, 202(4), p.109. Karimi, Z., 2021. Confusion Matrix. Encycl. Mach. Learn. Data Min., no. October, pp.260-260.

[11] Freiesleben, T. and Grote, T., 2023. Beyond generalization: a theory of robustness in machine learning. Synthese, 202(4), p.109.