Optimizing Imbalanced Data Classification: Under Sampling Algorithm Strategy with Classification Combination

Nauval Dwi Primadya[1], Adhitya Nugraha[2], Sahrul Yudha Fahrezi[3], Ardytha Luthfiarta[4]

Program Studi Teknik Informatika,
Fakultas Ilmu Komputer,
Universitas Dian Nuswantoro, Semarang
[1]primadya021@gmail.com, [2]adhitya@dsn.dinus.ac.id, [3]yudhafahrezi30@gmail.com,
[4]ardytha.luthfiarta@dsn.dinus.ac.id

Abstract

The security of Internet of Things devices is a factor that must be considered because device damage and data theft can occur. Internet of Things devices are very useful in various sectors, such as health, transportation, and industrial sectors. Attacks on Internet of Things devices increase every year. To overcome this, it is necessary to take a research approach with machine learning. The dataset used is CIC IoT Attacks 2023 from the University of New Brunswick. To be able to produce good data, it is necessary to do random under-sampling as a way to overcome data imbalance. Then, modeling is done using the KNN algorithm, Random Forest, Logistic Regression, Adaboost, And Perceptron. The result of this research is that random forest has the best accuracy result of 99.73%. From these results, it can be concluded that the random under-sampling technique can improve the accuracy of data imbalance.

Keyword: Random under-sampling, IoT Attacks 2023, combination algorithm

Abstrak

Keamanan perangkat internet of thing adalah faktor yang harus diperhatikan karena dapat terjadi kerusakan perangkat dan pencurian data. Dimana perangkat internet of thing sangat bermanfaat di berbagai sektor seperti sektor kesehatan, transportasi, dan industri. Penyerangan terhadap perangkat internet of thing setiap tahunnya meningkat. Untuk mengatasi hal tersebut perlu dilakukan penelitian pendekatan dengan machine learning. Dataset yang digunakan CIC IOT ATTACKS 2023 dari University of New Brunswick. Untuk dapat menghasilkan data yang baik maka perlu dilakukan random undersampling sebagai salah satu cara mengatasi data imbalance. Kemudian dilakukan pemodelan menggunakan algoritma KNN, Random Forest, Logistic Regression, Adaboost, dan Perceptron. Hasil penelitian ini adalah Random Forest memiliki hasil akurasi terbaik sebesar 99.73%. Dari hasil tersebut, dapat disimpulkan teknik random undersampling dapat meningkatkan akurasi dari ketidak seimbangan data.

Kata Kunci: Random under-sampling, IoT Attacks 2023, combination algorithm

Received: 28 Desember 2023 Revised: 28 Maret 2024 Accepted: 16 April 2024

1. Introduction

The Internet of Things is a concept that can connect electronic devices to the Internet network, allowing them to interact and share information. The Internet of Things has brought significant benefits in various sectors, such as health, transportation, and industry [1]. But in the course of time and technological sophistication, IoT security has become increasingly important. This is because it is always connected to the internet. The vulnerability of IoT devices can cause damage to the device and or theft of personal data [2].

One type of attack that often occurs on IoT devices is distributed denial of service or DDoS. The result of an attack from DDoS is to make the IoT device unresponsive and can even result in a complete shutdown [3]. To be able to protect IoT devices from various attacks, not only DDoS but datasets are also needed that can be used to train machine learning models to be able to detect these attacks.

In the search for datasets, there are often balanced datasets and imbalance datasets, it is very common and natural to find these two cases. A balanced dataset is a dataset where the class distribution has a balance. Meanwhile, an imbalanced dataset is a dataset where the class distribution is not balanced or evenly distributed. This can cause the classification process to lean towards the majority class rather than the minority class. Data imbalance is a common problem in machine learning, where the majority class will be larger than the minority class [4]. The impact of the data imbalance is that it will affect the performance of the classifier such as noise and insufficient number of training data observations. The emergence of data imbalance cases is one of the developments in machine learning.

The application of classification algorithms without considering the balance of data classes will produce good predictions for the majority class but ignore the minority class. The methods used to overcome data imbalance are divided into two, namely under sampling and oversampling. Under sampling is a technique to reduce the number of majority classes to be equal to the minority class. While oversampling is a technique to equalize the minority class with the majority [5], [6]. There are various ways to use these two methods; for example, in oversampling, there is random under sampling, smote, Adasyn, and so on. While in under sampling, there is random under sampling, cluster-based under sampling, near-miss, and so on [6].

A study conducted by Gagah Gumelar found that applying under sampling methods to unbalanced datasets with a combination of various classification algorithms can result in improved performance on all classification algorithms used [7]. In this study, they used four datasets that had different levels of imbalance and applied classification algorithms such as c45, Naïve Bayes, KNN, and SVM. Furthermore, they compared the results before and after applying the under-sampling method. The results showed that the use of the under-sampling method significantly improved the performance of all classification algorithms used [8].

In 2023, Fauzi Adi Rafrastara and his team conducted research. In this study, they utilized the random under sampling (RUS) method to overcome the imbalance of the dataset, and they used the random forest algorithm to classify files as good ware or malware [9]. The results showed that in comparison with other methods, random forest achieved the most optimal performance, with accuracy reaching 98.1%, recall reaching 98.0%, and specificity reaching 98.2% [9].

Rizki Fauziyah and her team conducted research. In this study, they applied the random over-under sampling (ROUS) method to overcome the imbalance of the dataset and used classification algorithms to predict student graduation. The findings of this study showed that the application of the ROUS method successfully improved the performance of the classification model by producing an increase in accuracy value [10].

Ilham Kurniawan and his team conducted research. In the study, they applied the random under sampling (RUS) method to overcome dataset imbalance and used a decision tree algorithm to predict forest fires. The results of this study indicated that the decision tree algorithm successfully achieved an accuracy rate of 94.52% [11].

2. Literature Study

2.1. Preprocessing

Preprocessing in machine learning is a data processing phase that is applied before the machine learning model uses the data. The purpose of preprocessing is to prepare the data so that it can be properly processed by machine learning models and produce accurate results. Some preprocessing methods that are often used in machine learning involve normalization, missing data elimination, categorical variable coding, and dimensionality reduction [12].

2.2. Random Under Sampling

The random under-sampling (RUS) method is a strategy used to address imbalances in a dataset by reducing the number of samples from the majority class. In this method, samples are randomly selected from the majority class so that their number becomes equal to the number of samples in the minority class. The application of rules is proven to be very efficient in overcoming class imbalance in datasets and is able to improve the performance of machine learning models [13].

2.3. Data Splitting

Data splitting, also known as data splitting, is a technique that generates two or more subsets of data by dividing a dataset into separate parts. Typically, this data splitting produces two subsets, where the first subset is used for data evaluation or testing, while the other subset is used for training the model. Data splitting is a very important aspect of data science, especially in data-driven model building. This technique plays a role in ensuring that the model created has sufficient accuracy and can be applied in advanced stages, such as in machine learning processes [14].

2.4. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is one of the machine learning algorithms used in classification and regression tasks [15]. This approach is done by finding a number of k closest data points to the data that you want to perform classification or regression on. It then selects the class or regression value that occurs most frequently among the k data points. KNN is well-known for its simplicity in implementation and is able to provide satisfactory results, especially on datasets that have a relatively small size [16].

$$dis = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \tag{1}$$

where x_1 is sample data, dis is distance, x_2 is test data, n is data dimension, and i is data variable.

2.5. Random Forest

Random forest is one of the machine learning methods used for classification and regression tasks [17]. It operates by combining multiple decision trees to improve accuracy and reduce overfitting problems. Random forest is quite popular due to its ease of implementation, its ability to cope with data imbalance, and its applicability for classification involving many classes (multiclass).

2.6. Logistic Regression

Logistic regression is one of the statistical techniques utilized to project the binary outcome of a dependent variable based on one or more independent variables. This approach has gained great popularity in the analysis of data involving categorical variables and is widely used in various disciplines, including economics, medicine, and social sciences [18].

$$Ln\left(\frac{p}{1-p}\right) = B_o + B_1X \tag{2}$$

$B_o = Constant$

$B_1 = The\ coefficient\ of\ each\ variable$

The value of p or chance (Y=1) can be found in the equation

$$p = \frac{e^{(B_O+B_1X)}}{(1+e^{(B_O+B_1X)}} \tag{3}$$

2.7. Adaboost

Adaboost, also known as adaptive boosting, is one of the ensemble learning methods used to improve the performance of machine learning models. This approach works by combining a number of machine learning models that have relatively weak performance into one model that has greater strength. Adaboost has become a popular choice due to its ease of implementation and its ability to be used in various types of problems, be it classification or regression tasks [19], [20].

$$H(x) = sign\ (\textstyle\sum_{t=1}^{T} a_t h_t(x)) \tag{4}$$

2.8. Perceptron

The Perceptron is a machine learning algorithm used to perform binary classification. It involves calculating the weights of each input feature and producing an output based on an activation function. Perceptron is very simple and easy to apply, although its limitation is that it can only be used in the case of linearly separable binary classification problems [21].

$$n = \textstyle\sum x_i\, w_i + b \tag{5}$$

2.9. Confusion Matrix

The confusion matrix is a machine learning model performance evaluation technique used to measure the model's precision in classification. This matrix describes the amount of data that has been correctly classified and the amount that the model has incorrectly classified. The confusion matrix consists of four components, namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP is the amount of positive data that has been correctly classified, FP is the amount of negative data that has been incorrectly classified as positive, TN is the amount of negative data that has been correctly classified, and FN is the amount of positive data that has been incorrectly classified as negative [22].

Table 1. Confusion Matrix

Predicted class	Actual class	
	+	-
+	True positive (TP)	False positive (FP)
-	False negative (FN)	True negative (TN)

To calculate accuracy, we use the following formula

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

Precision is a term that refers to the extent to which relevant items are selected compared to all selected items. In this context, precision reflects the extent to which the answer to an information request matches the request. The precision formula used to measure is as follows:

$$Precision = \frac{TP}{(TP+FP)} \tag{7}$$

Recall is a term that refers to the extent to which relevant items are selected in relation to the total number of relevant items available. The recall calculation is done using the following formula:

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

3. Research Method

The proposed research method is shown in Figure 1. To perform machine learning modeling, there are several stages as follows:

1. Phase 1 Data Understanding
   Collect data relevant to the object of research. The data set is based on data that has been taken from CIC IoT attacks 2023.
2. Phase 2 Data Processing
   Analyzing the imbalanced dataset, then random under sampling and cleaning the data so as to get balanced data. After the data is balanced, data splitting will be carried out, namely training models and testing data to test the model.
3. Phase 3 Modeling
   The data that has been prepared will be used for machine learning modeling using KNN, random forest, logistic regression, AdaBoost, and Perceptron so that a classification model is formed.
4. Phase 4 Evaluation
   Evaluate the model from the prediction results using training data using the classification report and confusion matrix.

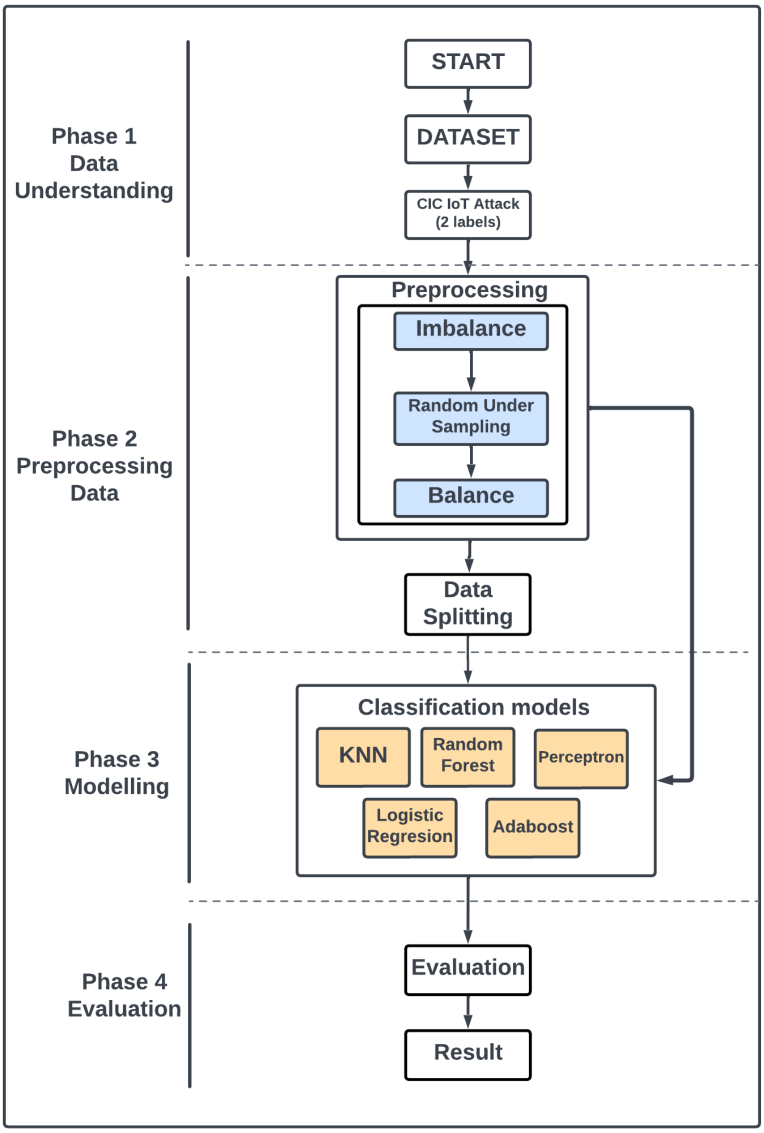

Figure 1. Proposed Research Method

4. Discussion and Result

4.1. Dataset

The dataset used in the research is a dataset compiled by the University of New Brunswick under the name CIC IoT Attacks 2023, which contains various types of attacks on IoT devices. The dataset is used as material for training machine learning. The data contained in the dataset is approximately 46.686.545 rows with 34 labels, which are used to train the model and test the model. Furthermore, from 34 labels, it is made into two labels with the label names "Attack" and "Benign" with the following comparison.

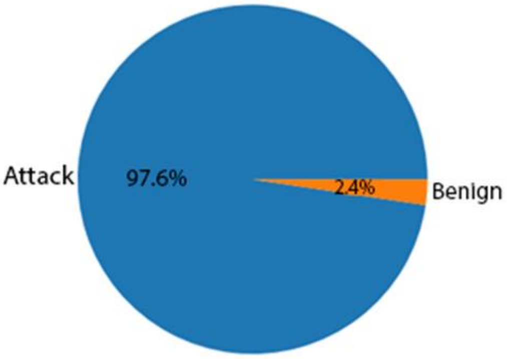

Figure 2. Dataset Before Balancing

4.2. Under-Sampling Using Random Under-Sampling

In the previous dataset results, the distribution of data distribution between classes is not balanced. Where the attack class has 97.6% data with a total of 45.588.350 and the benign class has 2.4% data with a total of 1.098.195 data. To overcome the data imbalance, a random under sampling model is used. Random under sampling will make the amount of majority data balanced with the majority data. In this study, the amount of attack class data will be equalized with the benign class. Before doing random under-sampling, the dataset will be divided into 2, namely the majority class and the minority class, which will then be processed to equalize the amount of data so that the unbalanced dataset becomes balanced and data cleaning is carried out to eliminate the possibility of duplicate data.

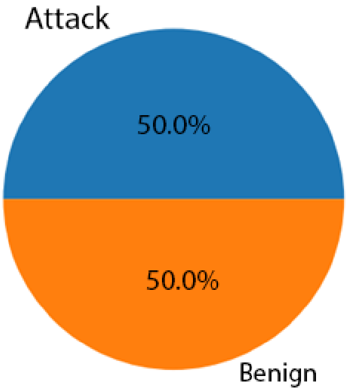

Figure 3. Dataset After Balancing

4.3. Modeling

After passing the random under-sampling process, the results of the dataset will be divided into 2, namely training data with an amount of 80% and 20% testing data. After that, modeling will be carried out with several algorithms, namely KNN, random forest, logistic regression, AdaBoost, and Perceptron.

Table 2. Comparison after and before using random under sampling

ML classifier	After Use Random Under sampling			Before Use Random Under sampling		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	99.52%	99.08%	99.94%	99.44%	94.75%	93.32%
Random Forest	99.73%	99.48%	99.99%	99.68%	96.53%	96.51%
Logistic Regression	93.43%	95.88%	91.34%	98.90%	86.31%	89.04%
Adaboost	99.61%	99.29%	99.92%	99.58%	96.56%	94.73%
Perceptron	86.38%	95.84%	80.48%	98.17%	82.54%	79.70%

After implementing Random Under sampling (RUS) technique, the data from the two classes, namely attack and benign, became balanced. Prior to RUS, the data for the attack class was imbalanced compared to the benign class. By applying RUS, the number of samples from both classes became equal, eliminating the imbalance that existed previously. This ensures that the model is trained on a dataset that is balanced proportionally between the attack and benign classes, which can enhance the performance and fairness of the model in classifying both types of samples.

Before implementing random under-sampling, model evaluation showed that most models had high accuracy, ranging from 98.17% to 99.68%. However, they exhibited low precision and recall for minority classes, indicating limitations in identifying samples from those classes.

After implementing random under-sampling, there was a significant improvement in precision and recall for minority classes across almost all models. Precision values increased to a range of 95.84% to 99.48%, while recall values increased to a range of 80.48% to 99.99%. The KNN and Random Forest models, in particular, demonstrated significant improvements in precision and recall for minority classes. Other models also showed considerable enhancements. These findings indicate that random under-sampling effectively enhances the models' ability to recognize and classify minority classes, reducing the risk of bias towards majority classes.

The choice of K values in the KNN algorithm, such as K=3 and K=5, has a significant impact on classification results. Different K values are selected to control model complexity and avoid overfitting or underfitting. Smaller K values, like K=3, result in complex decision boundaries that are sensitive to minor fluctuations, noise, and outliers in the data. On the other hand, larger K values, such as K=5, yield smoother and more stable decision boundaries, are more resilient to noise and outliers, and may result in more general and interpretable models due to considering a larger number of nearest neighbors.

The choice between K=3 and K=5 depends on the specific characteristics of the data and the complexity of the classification problem. Smaller K values can be useful when dealing with complex or changing patterns in the data. Larger K values can provide more stable and generalized results. Therefore, the selection of the appropriate K value should consider the trade-off between model complexity, robustness to noise, and interpretability of the results.

Overall, the evaluation results confirm that using random under-sampling techniques is an effective approach to address class imbalance in the dataset. Although it may lead to a decrease in accuracy, the substantial improvements in precision and recall for minority classes indicate that the model becomes more reliable in correctly classifying samples from all classes.

4.4. Evaluation

By using known as a classifier, the results obtained are 99.52% for accuracy, 99.08% for precision, and 99.94% for recall. These results are shown in Figure 4. The study used K values of 3 and 5, where K values of 3 had the highest value.

Interestingly, the random forest gets the highest accuracy result of 99.73%, with a precision of 99.48% and a recall of 99.99%. Where as many as 215898 are considered attacks and 219620 as benign, which can be seen in Figure 5.

The results of matrix logistic regression proved as many as 208089 as an attack and 199913 as benign by producing an accuracy rate of 93.43%.

Using Adaboost resulted in an accuracy of 99.61%, precision of 99.29%, and recall of 99.92%. These results can be seen from the confusion matrix results in Image 7, where 215493 is an attack, and 219472 is benign.

Finally, the classification using Perceptron resulted in 86.38% accuracy, 95.84% precision, and 80.48% recall.

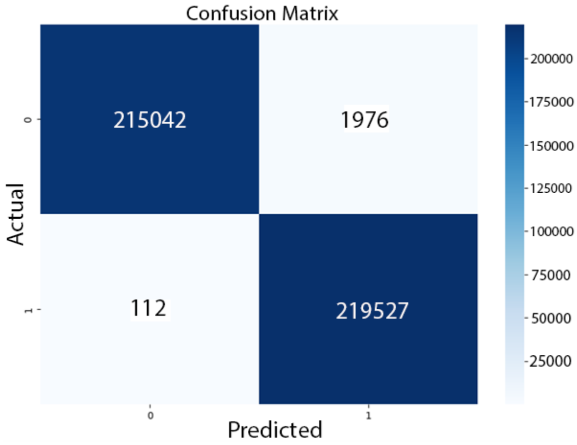

Figure 4. Confusion Matrix Algorithm KNN

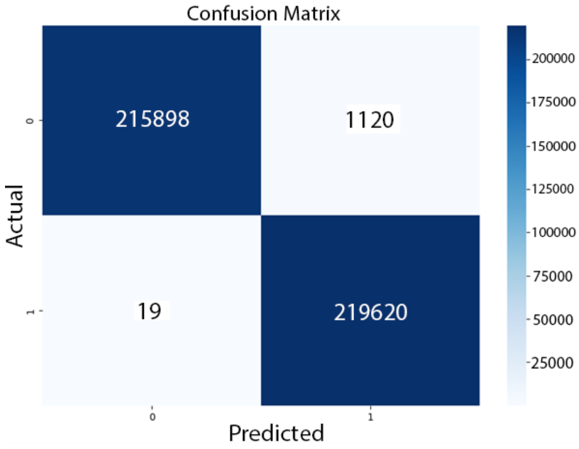

Figure 5. Confusion Matrix Algorithm Random Forest

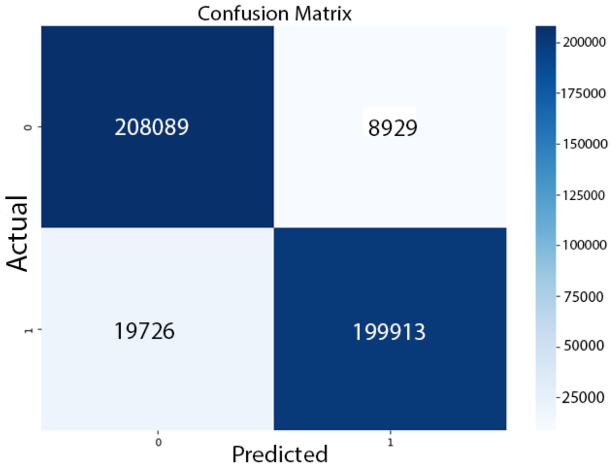

Figure 6. Confusion Matrix Algorithm Logistic Regression

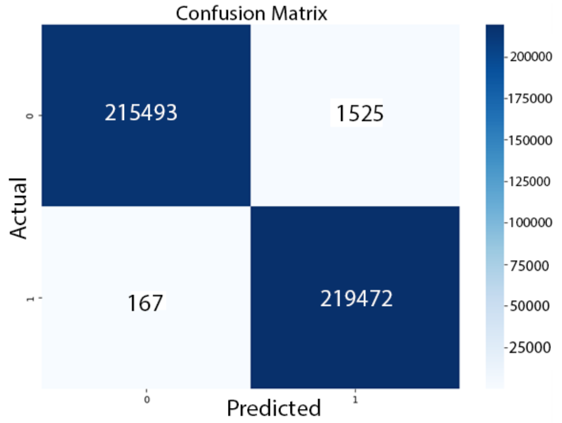

Figure 7. Confusion Matrix Algorithm Adaboost

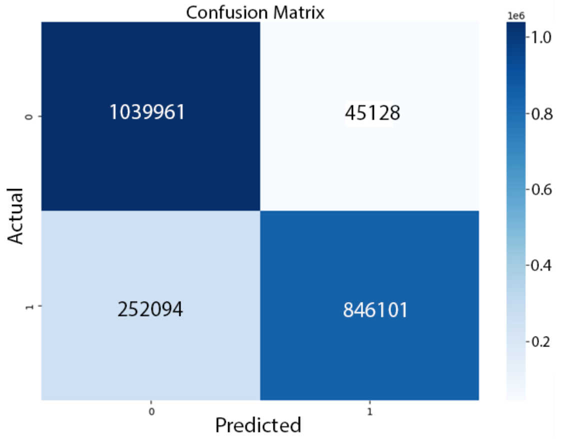

Figure 8. Confusion Matrix Algorithm Perceptron

The comparison test of resampling techniques is carried out by comparing datasets that have undergone resampling with datasets that do not use resampling techniques. The random under-sampling resampling technique is used to manage class imbalance in the data set. Based on the results of experiments and tests that have been carried out, the following conclusions can be drawn:

1. The application of resampling random under-sampling combined with classification algorithms can increase accuracy in the random forest algorithm by 0.05% and AdaBoost by 0.03%.
2. The application of resampling random under-sampling combined with the logistic regression algorithm decreased by 5.47%, and the perceptron algorithm decreased by 18.72%.

The cause of the decrease in accuracy is due to the imbalance of classes in a dataset. Using random under-sampling is one way to overcome the class imbalance. However, by doing random under-sampling, there is a random deletion of data and allows the loss of important data, which results in classification results. Therefore, a decrease in accuracy can occur after random under-sampling.

5. Conclusion

Experiments with the CIC IoT Attacks 2023 dataset show that the use of random under-sampling techniques significantly addresses class imbalance. In this process, the data distribution between the attack and benign classes is balanced by reducing the number of majority data so that it aligns with the number of minority data. However, a decrease in accuracy can occur in some classification algorithms, such as logistic regression and perceptron, with a decrease of 5.47% and 18.72% respectively. This suggests a compromise between achieving class balance and maintaining model accuracy.

In contrast, the experimental results show that the use of certain classification algorithms, such as random forest and AdaBoost, results in improved accuracy after applying random under-sampling. Although this accuracy improvement is small, it is statistically significant, indicating that applying resampling techniques, such as random under-sampling, is beneficial in improving model performance in special cases, especially in dealing with class imbalance. Therefore, the selection of appropriate resampling techniques and classification algorithms is crucial to achieve a balance between accuracy and handling class imbalance in data analysis.

Reference

[1] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, “CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment,” Sensors, vol. 23, no. 13, Jul. 2023, doi: 10.3390/s23135941.

[2] M. Zolanvari, M. A. Teixeira and R. Jain, "Effect of Imbalanced Datasets on Security of Industrial IoT Using Machine Learning," 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, USA, pp. 112-117, 2018, doi: 10.1109/ISI.2018.8587389.

[3] Al-Hadhrami, Y., and Hussain, F. K., “DDoS attacks in IoT networks: a comprehensive systematic literature review,” *World Wide Web*, vol. 24, no. 3, pp. 971-1001, 2021

[4] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” J Big Data, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00349-y.

[5] N. Noorhalim, A. Ali, and S. M. Shamsuddin, “Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE,” in Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017), Springer Singapore, 2019, pp. 19–30. doi: 10.1007/978-981-13-7279-7_3.

[6] A. Ilham, “Komparasi Algoritma Klasifikasi Dengan Pendekatan Level Data Untuk Menangani Data Kelas Tidak Seimbang,” Jurnal Ilmiah Ilmu Komputer, vol. 3, no. 1, 2017, [Online]. Available: http://ejournal.fikom-unasman.ac.id

[7] G. Gumelar, Norlaila, Q. Ain, Riza Marsuciati, S. A. Bambang, A. Sunyoto, and M. S. Mustafa, “Kombinasi Algoritma Sampling Dengan Algoritma Klasifikasi Untuk Meningkatkan Performa Klasifikasi Dataset Imbalance,” Prosiding SISFOTEK, vol. 5, no. 1, 250–255, 2021. Available: https://www.seminar.iaii.or.id/index.php/SISFOTEK/article/view/295

[8] A. Y. Triyanto and R. Kusumaningrum, “Implementasi Teknik Sampling untuk Mengatasi Imbalanced Data pada Penentuan Status Gizi Balita dengan Menggunakan Learning Vector Quantization Implementation of Sampling Techniques for Solving Imbalanced Data Problem in Determination of Toddler Nutritional Status using Learning Vector Quantization,” vol. 19, pp. 39–50, 2017.

[9] F. A. Rafrastara, C. Supriyanto, C. Paramita, Y. P. Astuti, and F. Ahmed, “Performance Improvement of Random Forest Algorithm for Malware Detection on Imbalanced

Dataset using Random Under-Sampling Method," vol. 8, no. 2, 2023, [Online]. Available: https://orangedatamining.com/

[10]E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," *Bianglala Informatika*, vol. 10, no. 1, pp. 42-47, 2022.

[11]I. Kurniawan, D.C.P. Buani, W.A. Abdussomad, and E. Fitriani, "Penerapan Teknik Random Undersampling untuk Mengatasi Imbalance Class dalam Prediksi Kebakaran Hutan Menggunakan Algoritma Decision Tree," *Forest*, vol. 14, 244, 2023.

[12]B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," JBASE - Journal of Business and Audit Information Systems, vol. 4, no. 2, Aug. 2021, doi: 10.30813/jbase.v4i2.3000.

[13]A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit," Jurnal Informatika, vol. 5, no. 2, 2018.

[14]A. Nurhopipah and U. Hasanah, "Dataset Splitting Techniques Comparison For Face Classification on CCTV Images," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 14, no. 4, p. 341, Oct. 2020, doi: 10.22146/ijccs.58092.

[15]R. Nuari, A. Apriliyani, J. Juwari, and K. Kusrini, "Implementasi Metode K-Nearest Neighbor (KNN) untuk Memprediksi Varietas Padi yang Cocok untuk Lahan Pertanian," *Jurnal Informa: Jurnal Penelitian dan Pengabdian Masyarakat*, vol. 4, no. 2, pp. 28-34, 2018.

[16]A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm for Public Sentiment Analysis of Online Learning," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 2, p. 121, Apr. 2021, doi: 10.22146/ijccs.65176.

[17]L. Fadilah, "Klasifikasi Random Forest pada data imbalanced", Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2018.

[18]W. A. Setyati, S. Sunaryo, A. Rezagama, A. K. Widodo, And M. F. A. Yulianto, "Penerapan Regresi Logistik Dalam Penentuan Faktor Yang Mempengaruhi Jumlah Wisatawan Ecotourism Desa Bedono," Jurnal Enggano, vol. 5, no. 1, pp. 11–22, Apr. 2020, doi: 10.31186/jenggano.5.1.11-22.

[19]A. Bisri and R. S. Wahono, "Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree," Journal of Intelligent Systems, vol. 1, no. 1, 2015, [Online]. Available: http://journal.ilmukomputer.org

[20]Z. K. S. Domas and R. Rakhmadi, "Peningkatan Performa Decision Tree dengan AdaBoost untuk Klasifikasi Kekurangtransparanan Informasi Anti-Korupsi," Applied Information System and Management (AISM), vol. 5, no. 2, pp. 75–82, Sep. 2022, doi: 10.15408/aism.v5i2.24887.

[21]S. Grania and T.M.S. Mulyana, "Penerapan Algoritma Perceptron Pada Jaringan Syaraf Tiruan Dalam Pembagian Jurusan," *Jurnal Teknologi Informasi*, vol. 11, no. 2, 2017

[22]I. Düntsch and G. Gediga, "Confusion matrices and rough set data analysis," *Journal of Physics: Conference Series*, vol. 1229, no. 1, 012055, 2019, doi: 10.1088/1742-6596/1229/1/012055.