ENHANCING FRAUD DETECTION PERFORMANCE IN E-COMMERCE PLATFORMS USING GRADIENT BOOSTING ALGORITHMS

Ardi Saputra[1], Fauzi Adi Rafrastara[2*], Wildanil Ghozi[3]
[1,2,3] Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207, Pendrikan Kidul,
Kec. Semarang Tengah, Kota Semarang, Jawa Tengah 50131
111202113790@mhs.dinus.ac.id[1], fauziadi@dsn.dinus.ac.id[2], wildanil.ghozi@dsn.dinus.ac.id[3]

Abstract - The rapid growth of e-commerce has attracted many users. However, as transaction volumes increase, so do cases of fraud. This not only causes financial losses for sellers but also threatens the trust that is so important in the e-commerce industry. Previous studies have used the Naïve Bayes and Multilayer Perceptron algorithms to detect fraud in e-commerce with accuracy percentages of 95.00% and 94.00%, respectively, without other assessment measures, including precision, recall, and F1-score. This research seeks to create a predictive model for the likelihood of online sales fraud by comparing Gradient Boosting, Neural Network, Random Forest, and Naïve Bayes models through feature extraction and feature scaling pre-processing, with 10-fold cross-validation. The dataset used was taken from the Kaggle platform. The features included in the dataset include buyer characteristics, products sold, transaction volume, devices used, and other fraud indicators. The study's findings demonstrate that the Gradient Boosting algorithm excels in detecting fraud risk with an accuracy rate of 95.30%, precision of 94.10%, recall of 95.30%, and an F1-score of 93.80%. These findings are anticipated to enhance the development of more efficient e-commerce security solutions.

Keywords - E-Commerce, Fraud, Gradient Boosting, Kaggle.

I. INTRODUCTION

Increased reliance on digital platforms has led to an increase in cybercrime and fraud[1], costing the global economy billions of dollars and endangering public safety [2]. These fraudulent attacks not only affect the reputation of platforms but also impact user experience and can result in the attrition of platform users [3]. Therefore, e-commerce platforms must establish resilient anti-fraud measures to mitigate financial losses [4]. The number of studies on e-commerce fraud detection is still very limited, and they mainly focus on identifying characteristics or quality [5]. This is evidenced by studies [6] and [7], which yielded accuracy rates of only 94% and 95%, even though fraud is subject to a “zero tolerance” policy. In this digital world, it is crucial to have a solid method for identifying fraud [8]. Technological advancements have significantly enhanced human life, particularly through machine learning automation, which reliably executes numerous computations on extensive data sets and accomplishes various classification tasks, including handwriting recognition, sign language interpretation, crime detection, and credit card fraud identification [9]. Machine learning algorithms and artificial intelligence empower organizations to analyze extensive datasets to detect patterns and abnormalities that may signify fraudulent conduct [10].

Various methodologies exist for identifying fraud in e-commerce transactions, as indicated in the features available in the dataset that will be used in this study. These features include account age days, shipping address, and billing address. Advancements in technology enable machine learning algorithms to examine transactions and detect patterns indicative of fraudulent activity [11].

This study aims to improve the model in previous studies, considering that fraud in e-commerce has serious impacts. To address this challenge, this study evaluates the performance of Gradient Boosting, Neural Network, Random Forest, and Naïve Bayes methods. By evaluating various methods, the most effective classification method can be found.

II. SIGNIFICANCE OF STUDY

A. *Literature Review*

In study[6], a risk prediction model for online sales fraud was developed using the Naïve Bayes algorithm. The dataset used was "Fraudulent E-commerce Transaction" downloaded from Kaggle. The dataset underwent pre-processing to remove data with values < 10 and > 60 in the "Customer Age" feature to avoid abnormal data distribution patterns. Additionally, a new feature called "Address Match" was added. This column contains the parameter value 1 for matching shipping and billing addresses and 0 for mismatched addresses. The study achieved an accuracy of 95,00% using the Naive Bayes algorithm. Next, [7] detected fraud in the "Fraudulent E-commerce" dataset downloaded from Kaggle using the Multilayer Perceptron (MLP) algorithm. Fraud detection in e-commerce used an MLP model optimized with cost-sensitive learning (CSL) to balance the unbalanced data. The MLP model with CSL achieved an accuracy of 94,00%.

Based on previous studies, it can be concluded that the Naïve Bayes and MLP algorithms produced accuracy rates of 95.00% and 94.00%, respectively, without other evaluation metrics such as precision, recall, and F1-score. This study aims to improve the previously developed model by evaluating the Gradient Boosting, Neural Network, Random Forest, and Naïve Bayes methods. Additionally, a more comprehensive evaluation matrix will be included compared to previous studies. Each model in this study was checked for overfitting by comparing the model results with the training data and testing data to produce more accurate results. With this approach, this study is expected to produce a more accurate fraud detection system for e-commerce, thereby reducing financial losses and enhancing user trust in e-commerce platforms.

B. *Research Method*

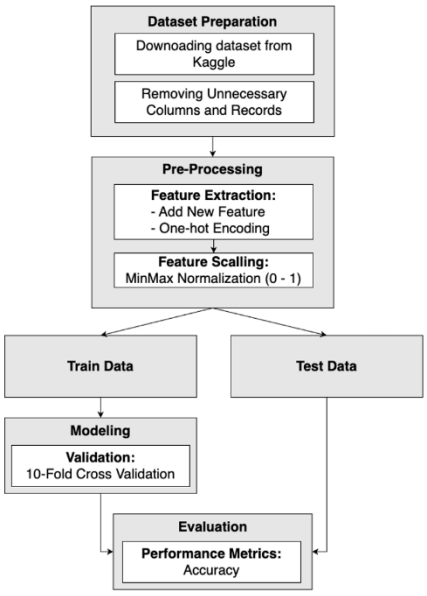

Figure 1. Research Stages

1. *Hardware and Software*

One of the factors that determine the smooth running and success of a research project is the supporting equipment. In computer science-based research, hardware and software play a very important role. Good software without the support of capable hardware will not be able to run optimally. Conversely, high-quality hardware that is not balanced with the right software will not be of much help. Therefore, hardware and software are essential in supporting the smoothness and success of a research project [12].

In this study, a personal computer with the following hardware specifications was prepared:

- Processor = Apple M2
- RAM = 8 GB
- SSD = 256 GB
-

Meanwhile, the software used in this study was Microsoft Excel and Orange (downloaded from https://orangedatamining.com/). Microsoft Excel was used to calculate the average evaluation matrix during five trials. Orange software was selected as the appropriate software for practicing data mining due to the graphical nature of the design process [13]. By using the Evaluate → Test and Score widget, the performance results of the Gradient Boosting model on both datasets were obtained, including accuracy, recall, precision, and F1-Score.

2. *Dataset Preparation*

This study uses Fraudulent E-commerce Transactions, which is public data downloaded from Kaggle. The details of the dataset are as follows:

- Dataset Name = Fraudulent E-Commerce Transactions
- Number of Data = 23634
- Number of Features = 16
- Missing Value = No
-

The target in the dataset is categorical, with 0 meaning no fraud and 1 meaning fraud. Fraudulent transactions account for 5% of the total data. The data is downloaded in a (.csv) file format. Next, to process the data, it is imported into the Orange application for further processing and analysis.

Advancements in technology enable machine learning algorithms to examine transactions and detect patterns indicative of fraudulent activity [14]. At this stage, the quantitative variable, namely customer age, is analyzed. It was found that this age data does not follow a normal distribution pattern because there are very small or very large values. For example, the lowest value (Q1) is 10, but there is age data that is less than that. Similarly, the highest value (Q3) is 60, but there are data points exceeding that number. Values outside the lower bound (Q1) and upper bound (Q3) are referred to as outliers. These outlier data can make the model less accurate. Therefore, the solution used is to remove the outlier data so that the model can function more effectively. For analysis purposes, not all columns will be used, so it would be better to remove irrelevant features to facilitate further analysis. The unused features are: “Transaction ID”, “Customer ID”, “IP Address”, “Customer Location”, and “Transaction Date”. Irrelevant features often affect the effectiveness of Machine Learning classification categorization [15]. The “Shipping Address” and “Billing Address” features are specialized to create a new feature called “Address Match”. A discussion of the “Address Match” feature will be included in the preprocessing section.

3. *Preprocessing*

The dataset contains Shipping Address and Billing Address fields that are used to record the address details associated with each transaction. For security analysis and potential fraud detection purposes, the two fields are compared to generate address matches that are entered into the Address Match field. This field will contain the parameter number 1 for matching shipping and billing addresses and 0 for mismatches. This address match column is used as one of the indicators in identifying transactions that are potentially suspicious or fraudulent. If there is a significant mismatch between the shipping address and the billing address, this can be one of the parameters for detecting transactions that fall into the risky or fraudulent category. Many datasets contain categorical features, which represent different categories or classes. However, most machine learning algorithms require numerical inputs, so label encoding is necessary. Label encoding allows machine learning models to convert categorical data into numerical inputs, consequently enhancing mathematical computations and analysis [16]. This process plays an important role when dealing with categorical data [17]. Each category is represented by a different integer value [18]. For example, 'paypal' could be 0, 'credit card' could be 1, and 'bank transfer' could be 2. This allows machine learning algorithms to process data while maintaining a meaningful representation of the categories. This study uses one-hot encoding to handle categorical data. In this process, a binary vector is created for each unique category, with all elements set to zero except for the one corresponding to the given observation category, which is set to one [19]. For example, for the variable “Product Category” with categories Electronic, Clothing, and Health, the corresponding one-hot encoding vector is:

- Electronic [1, 0, 0]
- Clothing [0, 1, 0]
- Health [0, 0, 1]

The subsequent stage in the pre-processing phase is Feature Scaling. Feature scaling is used to both numerical and categorical attributes. The objective is to standardize the feature range and mitigate bias. This research employed the Min-Max Normalization method. Min-max Normalization is a data scaling method that converts data into a defined range, usually between 0 and 1 [20]. Numeric features undergo Min-Max Normalization (0-1), whilst categorical characteristics are subjected to one-hot encoding and normalizing. Consequently, both numeric and categorical features maintain a uniform range from 0 to 1.

4. *Data Splitting*

This work employs distinct validation to partition data for model training and evaluation. A crucial element of a dataset for model evaluation and comprehension of its characteristics is the division of the dataset into training and testing subsets [21]. Training data is used for training and evaluating model results, while test samples are used to evaluate model performance on previously unseen data [22]. This experiment uses an 80:20 ratio for training and testing data. After the samples are split, the modeling process can begin.

5. *Modelling*

This study uses the Gradient Boosting, Random Forest, Neural Network, and Naïve Bayes classification algorithms. These algorithms were selected due to their popularity and effectiveness in data classification. The experiment was initially conducted without any feature selection. It was then repeated using the two feature selection methods mentioned in the previous sub-chapter. All four algorithms were implemented simultaneously in the first experiment. In the second and third experiments, only the best algorithm from the first experiment was used.

The Gradient Boosting model is a series of weak decision trees that are created one by one to form a strong learner [23]. A weak decision tree is initially created to forecast the output variable; thereafter, in the following iteration, another weak decision tree is developed to learn from the residual errors of the preceding weak decision tree [24]. Gradient boosting is beneficial for managing nonlinear relationships, accommodating extensive and diverse data sets, and progressively enhancing forecast accuracy through recurrent training [25]. Therefore, this makes gradient boosting a powerful tool for accurately predicting fraud in e-commerce transactions. The second algorithm is Neural Network, a type of artificial intelligence that mimics the function of the human brain [26]. Neural networks autonomously discern the most significant elements and relationships within data, so streamlining the learning process and enhancing result accuracy [27]. The neural network used in this experiment consists of 100 neurons. The algorithm uses the ReLU activation function and is configured with a maximum of 200 iterations.

The third method to be evaluated is Random Forest, which predicts outcomes by creating random decision trees and aggregating the predictions from each tree by majority voting (for classification) or averaging (for regression) to arrive at a final choice [28]. Random forests can be employed for both categorical response variables, known as "classification," and continuous responses, termed "regression" [29]. Although this algorithm can handle complex data with stable results, it requires relatively long processing time due to its high model complexity and lower interpretability [30]. The Naive Bayes algorithm is the final one to be evaluated. Naive Bayes is a probabilistic technique frequently employed for classification jobs. This algorithm functions under the presumption of independence among predictors or characteristics [31]. However, in many real-world cases, features are often interrelated, making this assumption often incorrect. As a result, the performance of Naive Bayes can be worse compared to other classification methods, especially for modern classification tasks [32].

6. *Evaluation*

Validation is necessary prior to evaluating the efficacy of each method. In this investigation, we employed 10-fold cross-validation (k=10). This technique mitigates overfitting in the model. Subsequent to validation, we assessed the model utilizing a Confusion Matrix. This matrix elucidates the prediction outcomes by computing the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as detailed below:

TABLE I
EVALUATION TERMS

Term	Explanation
TP (True Positive)	Predictions classified as correct based on actual data for fraud values
FP (False Positive)	Predictions classified as incorrect based on actual data for fraud values
TN (True Negative)	Predictions classified as correct based on actual data for non-fraudulent values
FN (False Negative)	Predictions classified as incorrect based on actual data for non-fraudulent values

These measures assess the efficacy of classification models, encompassing accuracy, recall, precision, and F-score. Our research emphasizes accuracy as a criterion for evaluation. Accuracy denotes the frequency with which the model makes true predictions [33]. The accuracy formula can be seen in Equation 1.

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

The F-score provides a balance between precision (Equation 2) and recall (Equation 3), making it very useful in situations with imbalanced class distributions, and therefore serves as a reliable overall performance metric [34]. The formula for calculating the F-score can be seen in Equation 4.

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

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

$$FScore = \frac{2 \times (precision \times recall)}{(precision + recall)} \quad (4)$$

III. RESULT AND DISCUSSION

Subsequent to the Data Preparation and Preprocessing phases, four algorithms Gradient Boosting, Neural Network, Random Forest, and Naïve Bayes were executed. The results of 10-fold cross-validation using the four algorithms can be seen in the following table.

TABLE II
MODELLING RESULT

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	95,30%	94,10%	95,30%	93,80%
Neural Network	95,22%	93,98%	95,20%	93,82%
Random Forest	95,10%	93,80%	95,10%	93,80%
Naïve Bayes	94,90%	90,00%	94,90%	92,40%

Figure 2 depicts the classification outcomes in a confusion matrix. The matrix indicates that the Gradient Boosting model misclassified data 1083 times out of 23368 trials. The number of non-fraudulent class predictions was 22099 correctly predicted transactions and 61 incorrectly predicted transactions. For the fraudulent class predictions, there were 186 correctly predicted transactions and 1022 incorrectly predicted transactions. Most classification errors occur in the form of false negatives, where fraudulent transactions are incorrectly predicted as non-fraudulent transactions. This is likely due to an imbalance in the dataset, where there are far more non-fraudulent transactions (22,160) than fraudulent transactions (1,208).

		Predicted 0	1	Σ
Actual	0	22099 (TN)	61 (FP)	**22160**
	1	1022 (FN)	186 (TP)	**1208**
	Σ	**23121**	**247**	**23368**

Figure 2. Matrix Evaluation

Next, the results of the Gradient Boosting model were compared with the test results using test data to test whether the model had overfitting or not. The test results using test data can be seen in Table 3. From Table 3, it is evident that the model results with test results using test data only differ by 0.2% in accuracy. This means that the model that has been created is not overfitting.

TABLE III
TEST ON TEST DATA

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	95,10%	94,00%	95,10%	93,82%
Neural Network	95,02%	93,76%	95,02%	93,72%
Random Forest	95,00%	93,80%	95,00%	93,86%
Naïve Bayes	94,60%	89,50%	94,60%	92,00%

The performance of the Gradient Boosting model was evaluated through preprocessing stages consisting of feature extraction (add new feature and one-hot encoding), and feature scaling compared to model [6] that underwent feature extraction preprocessing (add new feature) and model [7] that underwent feature scaling preprocessing. Both previous studies used the same dataset. The results can be seen in Table 4. Thus, the method we propose outperforms the two existing models.

TABLE IV
COMPARISON WITH STATE OF THE ART

Algorithm	Akurasi
Gradient Boost	**95,30%**
Naïve Bayes [12]	95,00%
MLP [13]	94,00%

IV. CONCLUSION

This study evaluates the performance of Gradient Boosting, Neural Network, Random Forest, and Naïve Bayes algorithms for detecting fraudulent transactions in e-commerce. After undergoing preprocessing stages that included feature extraction and feature scaling, the Gradient Boosting algorithm showed the best performance with an accuracy of 95.30%. These results indicate that Gradient Boosting is highly effective for detecting fraud compared to previous studies using the same dataset. This study contributes further by employing a more comprehensive preprocessing approach and evaluating performance using various metrics, thereby providing a more comprehensive understanding of the model's capabilities. Practically, the proposed model can be utilized by e-commerce platforms to enhance fraud detection systems, particularly in reducing prediction errors such as false positives and false negatives. For future research, it is recommended to explore handling data imbalance, such as resampling techniques or classification threshold adjustment, to improve the model's sensitivity to rare fraud cases.

REFERENCES

[1] Tedyyana, Agus, et al. "Enhance Telecommunication Security Through the Integration of Support Vector Machines." *International Journal of Advanced Computer Science & Applications* 15.3 (2024).

[2] A. Mutemi and F. Bacao, "E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review," Big Data Mining and Analytics, vol. 7, no. 2, pp. 419–444, Jun. 2024, doi: 10.26599/BDMA.2023.9020023.

[3] J. Yu et al., "Group-based Fraud Detection Network on e-Commerce Platforms," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2023, pp. 5463–5475. doi: 10.1145/3580305.3599836.

[4] Z. Wang, Q. Wu, B. Zheng, J. Wang, K. Huang, and Y. Shi, "Sequence As Genes: An User Behavior Modeling Framework for Fraud Transaction Detection in E-commerce," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2023, pp. 5194–5203. doi: 10.1145/3580305.3599905.

[5] S. Ray, "Fraud Detection in E-Commerce Using Machine Learning," BOHR International Journal of Advances in Management Research, vol. 2022, no. 1, pp. 7–14, 2022, doi: 10.54646/bijamr.002.

[6] L. Ayu Diah Pasha and Z. Azis, "Predicting The Risk of Online Sales Fraud with The Naïve Bayes Approach on Facebook Social Media," 2025, doi: 10.56211/hanif.v2i2.41.

[7] W. Priatna, H. D. Purnomo, A. Iriani, I. Sembiring, and T. Wellem, "Optimizing Multilayer Perceptron with Cost-Sensitive Learning for Addressing Class Imbalance in Credit Card Fraud Detection," Jurnal RESTI, vol. 8, no. 4, pp. 563–570, Aug. 2024, doi: 10.29207/resti.v8i4.5917.

[8] Y. W. Bhowte, A. Roy, K. B. Raj, M. Sharma, K. Devi, and P. Lathasoundarraj, "Advanced Fraud Detection Using Machine Learning Techniques in Accounting and Finance Sector," in Proceedings of 9th International Conference on Science, Technology, Engineering and Mathematics: The Role of Emerging Technologies in Digital Transformation, ICONSTEM 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICONSTEM60960.2024.10568756.

[9] B. Karunachandra, N. Putera, S. R. Wijaya, D. Suryani, J. Wesley, and Y. Purnama, "On the benefits of machine learning classification in cashback fraud detection," in Procedia Computer Science, Elsevier B.V., 2022, pp. 364–369. doi: 10.1016/j.procs.2022.12.147.

[10] S. R. Byrapu Reddy, P. Kanagala, P. Ravichandran, D. R. Pulimamidi, P. V. Sivarambabu, and N. S. A. Polireddi, "Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics," Measurement: Sensors, vol. 33, p. 101138, Jun. 2024, doi: 10.1016/j.measen.2024.101138.

[11] M. Gölyeri, S. Çelik, F. Bozyiğit, and D. Kılınç, Fraud Detection on E-commerce Transactions Using Machine Learning Techniques, vol. 3, no. 1. 2023. [Online]. Available: https://www.boyner.com.tr/

[12] D. Eko Waluyo et al., "Implementasi Algoritma Regresi pada Machine Learning untuk Prediksi Indeks Harga Saham Gabungan," Universitas Dian Nuswantoro, Semarang Jln. Imam Bonjol, vol. 9, no. 1, 2024, doi: 10.30591/jpit.v9i1.6105.

[13] Z. Dobesova, "Evaluation of Orange data mining software and examples for lecturing machine learning tasks in geoinformatics," Computer Applications in Engineering Education, vol. 32, no. 4, Jul. 2024, doi: 10.1002/cae.22735.

[14] F. Habibzadeh, "Data Distribution: Normal or Abnormal?," J Korean Med Sci, vol. 39, no. 3, 2024, doi: 10.3346/jkms.2024.39.e35.

[15] G. S. Rao and G. Muneeswari, "A Review: Machine Learning and Data Mining Approaches for Cardiovascular Disease Diagnosis and Prediction," EAI Endorsed Trans Pervasive Health Technol, vol. 10, 2024, doi: 10.4108/eetpht.10.5411.

[16] Md. A. Talukder, R. Hossen, M. A. Uddin, M. N. Uddin, and U. K. Acharjee, "Securing Transactions: A Hybrid Dependable Ensemble Machine Learning Model using IHT-LR and Grid Search," Feb. 2024, doi: 10.1186/s42400-024-00221-z.

[17] F. A. Rafrastara, G. F. Shidik, W. Ghozi, N. Rijati, and O. Setiono, "Tree-based Ensemble Algorithms and Feature Selection Method for Intelligent Distributed Denial of Service Attack Detection," Journal of Mobile Multimedia, vol. 14, no. 1, pp. 1–24, 2025, doi: 10.13052/jcsm2245-1439.1411.

[18] E. Manai, M. Mejri, and J. Fattahi, "Impact of Feature Encoding on Malware Classification Explainability," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.05614

[19] J. I. Samuels, “One-Hot Encoding and Two-Hot Encoding: An Introduction,” 2022, doi: 10.13140/RG.2.2.21459.76327.

[20] A. Pranolo et al., “Enhanced Multivariate Time Series Analysis Using LSTM: A Comparative Study of Min-Max and Z-Score Normalization Techniques,” ILKOM Jurnal Ilmiah, vol. 16, no. 2, pp. 210–220, 2024, doi: 10.33096/ilkom.v16i2.2333.210-220.

[21] Himanshu Sinha, “An examination of machine learning-based credit card fraud detection systems,” International Journal of Science and Research Archive, vol. 12, no. 2, pp. 2282–2284, Aug. 2024, doi: 10.30574/ijsra.2024.12.2.1456.

[22] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, “Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach,” Big Data and Cognitive Computing, vol. 8, no. 1, Jan. 2024, doi: 10.3390/bdcc8010006.

[23] P. Messer and T. Schmid, “Gradient Boosting for Hierarchical Data in Small Area Estimation,” Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.04256

[24] Tedyyana, Agus, Osman Ghazali, and Onno Purbo. "Model Design of Intrusion Detection System on Web Server Using Machine Learning Based." *Proceedings of the 11th International Applied Business and Engineering Conference, ABEC 2023, September 21st, 2023, Bengkalis, Riau, Indonesia*. 2024.

[25] W. H. Alawee, L. A. Al-Haddad, A. Basem, D. J. Jasim, H. S. Majdi, and A. J. Sultan, “Forecasting sustainable water production in convex tubular solar stills using gradient boosting analysis,” Desalination Water Treat, vol. 318, Apr. 2024, doi: 10.1016/j.dwt.2024.100344.

[26] J. Bintoro, F. Adi Rafrastara, I. Aulia Latifah, W. Ghozi, and W. Yassin, “OPTIMIZING ANDROID MALWARE DETECTION USING NEURAL NETWORKS AND FEATURE SELECTION METHOD,” Jurnal Teknik Informatika (JUTIF), vol. 5, no. 6, pp. 1663–1672, 2024, doi: 10.52436/1.jutif.2024.5.6.3898.

[27] A. Kozlova, V. Kukartsev, V. Melnikov, G. Kovalev, and A. Stashkevich, “Finding dependencies in the corporate environment using data mining,” in E3S Web of Conferences, EDP Sciences, Oct. 2023. doi: 10.1051/e3sconf/202343105032.

[28] M. Alden Nayef Anargya, W. Ghozi, and F. Adi Rafrastara, “Optimizing IoV Attack Detection using Random Under Sampling Techniques,” Jurnal Informatika: Jurnal pengembangan IT, vol. 10, no. 1, 2025, doi: 10.30591/jpit.v10i1.8034.

[29] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” Babylonian Journal of Machine Learning, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[30] K. Faturrahman and A. Sucipto, “OPTIMASI RANDOM FOREST DENGAN TEKNIK PRUNING UNTUK PREDIKSI DATA NASABAH BMT AL-HIKMAH PERMATA,” Jurnal Informatika Teknologi dan Sains, Aug. 2024, doi: 10.51401/jinteks.v6i3.4715.

[31] F. Torres Cruz, “Prediction of Depression Level in University Students through a Naive Bayes based Machine Learning Model,” 2023. [Online]. Available: https://arxiv.org/abs/2307.14371

[32] T. Ige, C. Kiekintveld, A. Piplai, A. Waggler, O. Kolade, and B. H. Matti, “An investigation into the performances of the Current state-of-the-art Naive Bayes, Non-Bayesian and Deep Learning Based Classifier for Phishing Detection: A Survey,” Nov. 2024

[33] R. Rakhmat Sani, F. Adi Rafrastara, and W. Ghozi, “Kinetik: Game Technology, Information System,” Computer Network, Computing, Electronics, and Control Journal, vol. 4, no. 3, pp. 47–52, 2019, doi: 10.22219/kinetik.v10i1.2051.

[34] C. W. Lee, M. W. Fu, C. C. Wang, and M. I. Azis, “Evaluating Machine Learning Algorithms for Financial Fraud Detection: Insights from Indonesia,” Mathematics, vol. 13, no. 4, Feb. 2025, doi: 10.3390/math13040600.