Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Comparison of Random Forest and Support Vector Machine Algorithms in Behavioral Data Outlier Detection Customer Purchases

Louis[1]*, Octara Pribadi[2] , Feriani Astuti Tarigan[3]

[1,2]Informatics Engineering, STMIK Time, Medan, Indonesia
[3]Information Systems, STMIK Time, Medan, Indonesia
valentinlouis77@gmail.com[1], octarapribadi@gmail.com[2], ferianiastuti@stmik-time.ac.id[3]*

Abstract

In the digital era, customer purchasing behavior data has become a valuable asset for retail companies to understand consumer preferences and improve marketing strategies. However, analyzing this data often encounters challenges due to the presence of outliers—data points that deviate from the general pattern and can affect the accuracy of the analysis. This study aims to compare the effectiveness of two machine learning algorithms, namely Random Forest and Support Vector Machine (SVM), in detecting outliers within complex and non-linear customer purchasing behavior data. The dataset used in this study was obtained from the Kaggle platform, consisting of 2,240 records, and the comparison was implemented through a web-based system. The results show that the Random Forest algorithm detected 112 outliers, while SVM detected 126 outliers, with 60 of them identified by both algorithms. Random Forest demonstrated advantages in terms of efficiency and ease of implementation, whereas SVM, although more sensitive in detecting anomalies, required more complex parameter tuning. Therefore, within the context of the developed system, Random Forest is recommended as a more efficient algorithm for detecting outliers in customer purchasing behavior data.

***Keywords**: Outliers, Customer Purchasing Behavior Data, Random Forest, Support Vector Machine (SVM), Machine Learning*

1. Introduction

In the increasingly advanced digital era, data has become one of the most valuable assets for companies in various sectors, including the retail sector [1]. One of the most important data in this sector is customer purchasing behavior data. Customer purchasing behavior data is data that records various consumer shopping activities or patterns, such as purchase frequency, types of products purchased, purchase times, spending amounts, and brand preferences. This data is usually collected through Point of Sale (POS) systems, loyalty cards, online shopping applications, or customer surveys. By analyzing this data, companies can understand their customers' preferences and needs more deeply [2]. With proper analysis, this data can be used to optimize marketing strategies, improve customer experience, and ultimately increase company revenue [3]. However, purchasing behavior data often has its own challenges, one of which is the presence of outlier data or data that deviates from the general pattern. This outlier data can affect the results of the analysis and make the conclusions drawn less accurate [4].

Detecting outliers in customer purchasing behavior data is crucial because outliers can indicate anomalies such as fraud, sudden changes in customer preferences, or even errors in data collection. However, the main challenge in detecting these outliers is the complexity of the data itself, which is often non-linear, has a high dimension, and does not follow a normal distribution. Conventional outlier detection methods, such as z-score, IQR (Interquartile Range), or linear regression, often assume a simple and linear data distribution, making them less effective in handling data with complex patterns. As a result, these methods can miss important outliers or even identify normal data as outliers, which can ultimately reduce the accuracy of analysis and decision-making [5]. Therefore, a more sophisticated approach is needed, namely machine learning algorithms, to accurately detect outliers in complex purchasing behavior data.

To address the problem of outlier detection in customer purchasing behavior data, this study compares the Random Forest and Support Vector Machine (SVM) algorithms. Random Forest, an ensemble learning method, is known for its ability to handle high-dimensional and non-linearly complex data [6][7]. On the other hand, SVM is known for its ability to distinguish linearly inseparable data using hyperplanes [8]. By conducting this comparison, it is hoped that the most effective and efficient algorithm for detecting outliers in customer purchasing behavior data can be found, so that companies can make more informed decisions based on accurate analysis results.

2. Literature Review

2.1. Customer Purchasing Behavior

Customer purchasing behavior is a series of processes and activities that occur when consumers, whether individuals, groups, or organizations, are involved in searching for, selecting, purchasing, using, and evaluating products, services, ideas, or experiences. The main goal is to satisfy needs and desires, which are carefully considered in making purchasing decisions [9]. However, in collecting and analyzing behavioral data, outliers are often found that need to be identified so that the analysis results remain accurate and reliable.

2.2. Outlier

Outlier data is data that has its own characteristics and is different from other research and has extreme values for a variable [10].

Outliers can be classified based on their characteristics [13]:
1. Univariate Outliers**:** Outlier values on a single variable, usually detected by boxplot or distribution analysis.
2. Multivariate Outliers**:** Occurs when several variables show unusual patterns simultaneously, is difficult to detect and is usually done using PCA or Mahalanobis distance.
3. Contextual Outliers**:** Depends on the context of the data, for example, large transactions that are normal during discounts but suspicious at other times.
4. Global Outliers**:** Data that differs significantly from the entire dataset without considering context, easily detected with basic statistics.

This study focuses on Multivariate Outliers **,** which are visualized through PCA to show deviant customer purchasing behavior data.

2.3. Random Forest

Random Forest is an ensemble method consisting of a collection of decision trees used to classify data into a class. The initial step in determining a decision tree is to calculate the entropy and information gain values [11].

There are two formulas for the Random Forest algorithm in forming a scientific tree, including [11]:
1. Entropy.
   Entropy approach as a determinant of attribute impurity. The calculation of the entropy value can be seen in the following equation:

   $$Entropy(S) = \sum_{i=1}^{n} - p_i log_2\, p_i \qquad (1)$$

   Information:
   S : Collect nan dataset
   n : Number of classes
   p_i : Proportion of S_i te to S

2. Information Gain.
   Information gain is a measure of how much information is gained from a node-splitting process in a decision tree. This value plays a crucial role in determining the best attribute to split, especially at the top node and split nodes. Meanwhile, the calculation of the Gini index continues as long as the Gini value has not reached zero, and it stops when the value indicates that there is no longer any impurity in the data.

   $$Information\ Gain\ (A) = Entropy(S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si \qquad (2)$$

   Information:
   A : Attribute t
   S : Dataset set
   $|S_i|$: Number i values
   |S| : The amount of data

2.4. Support Vector Machine

Support Vector Machine (SVM) was first introduced in 1992 by Vapnik with Boser and Guyon. The basic concept of SVM is to use a linear classifier, which was then developed to solve non-linear problems by applying kernel tricks in high-dimensional space [12]. The SVM algorithm model is one of the algorithms of the classification method, which works by finding a line (hyperplane) to separate two groups of data [12].

A hyperplane is the best dividing line between two classes. To find a hyperplane, you can find the margin of the hyperplane and then find the maximum point. The margin is the distance between the closest data points between two different classes, which is called the support vector. The solid line in Figure 1-b shows the best hyperplane, because it lies directly between the two classes, while the support vector is symbolized by the red and yellow dots inside the black circle [12].

The SVM linear classification hyperplane is denoted:

$$\mathrm{f(x) = w, x + b = 0} \qquad (3)$$

From the equation above, we get class inequality +1 (negative)

$$\mathrm{w.x + b \leq\ +1} \qquad (4)$$

Class -1 inequality:

$$w.x + b \geq -1 \tag{5}$$

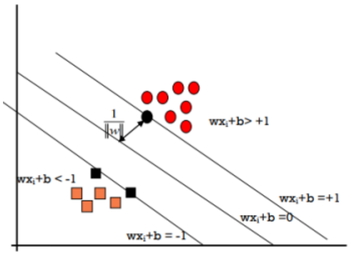


Fig. 1: Find the Optimal Separation Function for Linearly Separable Objects

w is the normal plane and b is the position of the plane relative to the center coordinate. By optimizing the distance value between the hyperplane and the next point, the largest margin can be found, namely 1 / ‖w‖. This can be formulated as a quadratic programming (QP) problem where the minimum point of equation (6) is given by the constraints of equation (7).

$$\min \frac{1}{2} || w ||^2 = \min \frac{1}{2}(w1^2 + w2^2) \tag{6}$$

$$y_i (w x_i + b) \geq 1, \; i = 1,2,3 \ldots, N \tag{7}$$

3. Research Methods

The stages of the research method in this study are shown in Figure 2.

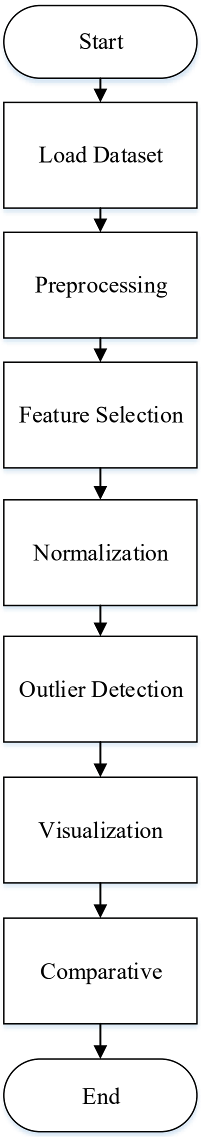


Fig. 2: Research Method Flowchart

3.1. Load Dataset

The initial stage was to collect a dataset, which was taken from Kaggle, specifically Consumer Buying Behavior Analysis, with 2,240 data points. Assume five data points were taken to analyze the manual calculations for *outlier detection* in the dataset, as can be seen in Tables 1, 2, 3, and 4.

Table 1: Example of a Dataset with the Attributes Year_Birth, Education, Martial_Status, Income, Kidhome, and Teenhome

ID	Year_Birth	Education	Marital Status	Income	Kidhome	Teenhome
5524	1957	Graduation	Single	58138	0	0
2174	1954	Graduation	Single	46344	1	1
4141	1965	Graduation	Together	71613	0	0
6182	1984	Graduation	Together	26646	1	0
5324	1981	PhD	Married	58293	1	0

Table 2: Example Dataset With Attributes Dt_Customer, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, and MntSweetProducts

Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
04-09-2012	58	635	88	546	172	88
03-08-2014	38	11	1	6	2	1
08-21-2013	26	426	49	127	111	21
02-10-2014	26	11	4	20	10	3
01-19-2014	94	173	43	118	46	27

Table 3: Example Dataset with Attributes MntGoldProds, NumDealsPurchases, NumWebPurchases, NumCatalogPurchase, NumStorePurchase, NumWebVisitsMonth, and AcceptedCmp3

MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3
88	3	8	10	4	7	0
6	2	1	1	2	5	0
42	1	8	2	10	4	0
5	2	2	0	4	6	0
15	5	5	3	6	5	0

Table 4: Example of a dataset with the attributes AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2, Complain, Z_CostContact, Z_Revenue, and Response

AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
0	0	0	0	0	3	11	1
0	0	0	0	0	3	11	0
0	0	0	0	0	3	11	0
0	0	0	0	0	3	11	0
0	0	0	0	0	3	11	0

3.2. Preprocessing

The preprocessing stages are divided into 2, namely eliminating missing values and labeling encoding on categorical columns.

3.3. Feature Selection

At this stage, feature selection is performed from the previously processed data. The goal of feature selection is to identify the most relevant and informative variables that represent customer purchasing behavior. The selected features are believed to significantly influence purchasing patterns and marketing campaign acceptance.

The selection of these features aims to provide comprehensive information about customers' economic backgrounds, their interactions with the company, spending patterns across product types, purchasing channel preferences, and responses to marketing efforts. This selected set of features will form the basis for the next stage, namely normalization and *outlier detection* .

3.4. Normalization

The normalization calculation in this study was carried out using StandardScaler , which is one of the normalization methods commonly used in machine learning practice to change the scale of features so that they have a distribution with a mean of 0 and a standard deviation of 1. This method works by calculating the difference between the original value of each data and the average of its column, then divided by the standard deviation of that column [13]. This normalization is important to ensure that each feature has a balanced contribution in the model learning process, especially in algorithms that are sensitive to data scale such as Support Vector Machine (SVM) and other distance-based algorithms. With StandardScaler, the data becomes more normally distributed, which can help improve the accuracy and stability of predictive models.

3.5. Outlier Detection

The next stage in data processing is outlier detection, which identifies data that deviates from the general pattern. This detection is crucial to ensure that extreme data that could negatively impact analysis results can be handled appropriately. Unrecognized outliers can reduce model accuracy, especially in complex and non-linear purchasing behavior datasets. At this stage, machine learning algorithms such as Random Forest and Support Vector Machine (SVM) are used to detect outliers based on preprocessed features.

3.6. Visualization

The visualization will display a scatter plot of the two principal components of PCA, with dots colored based on the outlier labels predicted by the Isolation Forest-based Random Forest and One-Class SVM.

3.7. Visualization

The comparative stage in this study aims to compare the capabilities of two machine learning algorithms, namely Isolation Forest-based Random Forest and One-Class SVM , in detecting outliers in customer purchasing behavior data. This process is carried out by identifying the number of data points detected as outliers by each algorithm and evaluating the similarity of detection results between the two.

4. Results

The research resulted in the development of a website for comparative analysis of the Random Forest and Support Vector Machine algorithms in detecting outliers in customer purchasing behavior. The following is a summary of the website's results:

1. Home/Login Page.
   The home/login page is the first interface a user sees when accessing the system. This design consists of form elements in the form of email and password text boxes, as seen in Figure 3.

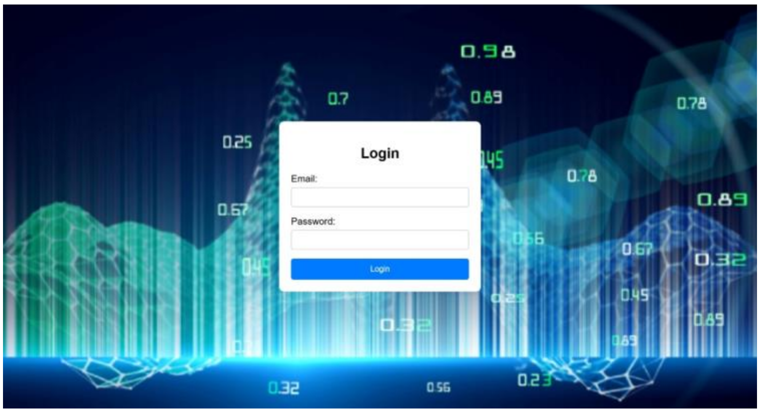
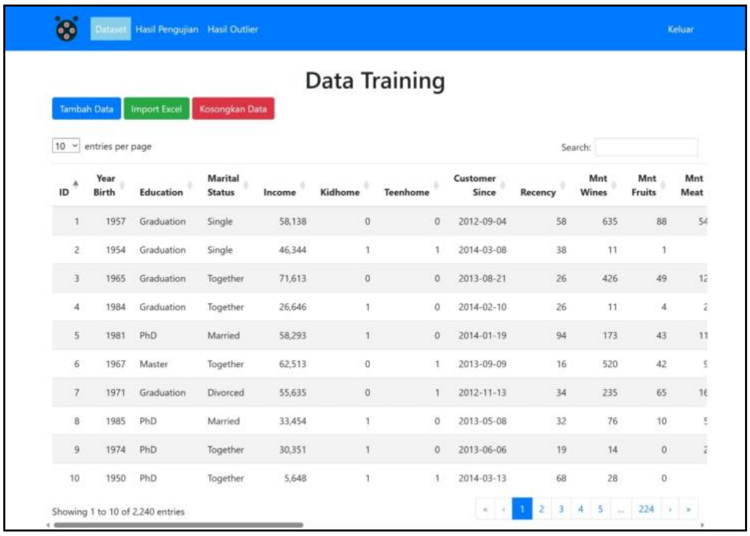


Fig. 3: Home/Login Page

2. Manage Dataset Page.
   The Manage Datasets page contains a *list of* available datasets. The Manage Datasets page can be seen in Figure 4.



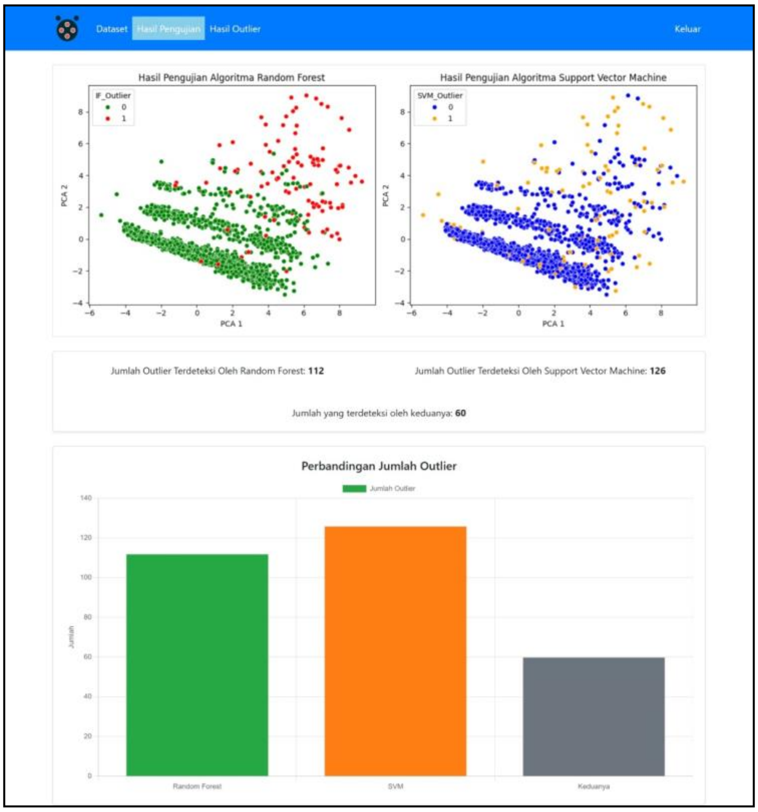
Data Training

ID	Year Birth	Education	Marital Status	Income	Kidhome	Teenhome	Customer Since	Recency	Mnt Wines	Mnt Fruits	Mnt Meat
1	1957	Graduation	Single	58,138	0	0	2012-09-04	58	635	88	54
2	1954	Graduation	Single	46,344	1	1	2014-03-08	38	11	1	
3	1965	Graduation	Together	71,613	0	0	2013-08-21	26	426	49	12
4	1984	Graduation	Together	26,646	1	0	2014-02-10	26	11	4	2
5	1981	PhD	Married	58,293	1	0	2014-01-19	94	173	43	11
6	1967	Master	Together	62,513	0	1	2013-09-09	16	520	42	9
7	1971	Graduation	Divorced	55,635	0	1	2012-11-13	34	235	65	16
8	1985	PhD	Married	33,454	1	0	2013-05-08	32	76	10	5
9	1974	PhD	Together	30,351	1	0	2013-06-06	19	14	0	2
10	1950	PhD	Together	5,648	1	1	2014-03-13	68	28	0	

Showing 1 to 10 of 2,240 entries

Fig. 4: Manage Dataset Page

3. Comparison Results Page.
   The comparison results page displays the results of testing the Random Forest and Support Vector Machine algorithms to detect outliers in customer purchasing behavior data. The comparison results page can be seen in Figure 5.



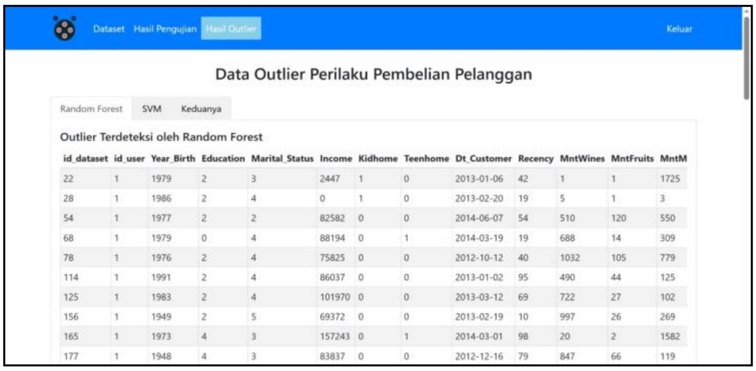
Jumlah Outlier Terdeteksi Oleh Random Forest: **112**

Jumlah Outlier Terdeteksi Oleh Support Vector Machine: **126**

Jumlah yang terdeteksi oleh keduanya: **60**

Fig. 5: Comparison Results Page

4. Outlier Results Page.
   The outlier results page contains the results of generating outlier data from the customer purchasing behavior dataset. The outlier detection results page using the Random Forest algorithm can be seen in Figure 6.



Data Outlier Perilaku Pembelian Pelanggan

Outlier Terdeteksi oleh Random Forest

id_dataset	id_user	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntM
22	1	1979	2	3	2447	1	0	2013-01-06	42	1	1	1725
28	1	1986	2	4	0	1	0	2013-02-20	19	5	1	3
54	1	1977	2	2	82582	0	0	2014-06-07	54	510	120	550
68	1	1979	0	4	88194	0	1	2014-03-19	19	688	14	309
78	1	1976	2	4	75825	0	0	2012-10-12	40	1032	105	779
114	1	1991	2	4	86037	0	0	2013-01-02	95	490	44	125
125	1	1983	2	4	101970	0	0	2013-03-12	69	722	27	102
156	1	1949	2	5	69372	0	0	2013-02-19	10	997	26	269
165	1	1973	4	3	157243	0	1	2014-03-01	98	20	2	1582
177	1	1948	4	3	83837	0	0	2012-12-16	79	847	66	119

Fig. 6: Outlier Detection Results Page with Random Forest Algorithm

5. Conclusion

Based on the results of research and the implementation of a website-based comparison system for the Random Forest and Support Vector Machine (SVM) algorithms in detecting outliers in customer purchasing behavior data, several important conclusions were obtained. The website that was built was able to demonstrate that both algorithms were effective in identifying outliers in complex and non-linear data. The test results showed that the Random Forest algorithm successfully detected 112 outliers, while the SVM algorithm detected 126 outliers. Of these, there were 60 data points that were identified as outliers by both algorithms, indicating similarities in some detection results, but there were still significant differences due to the different approaches of each method in detecting anomalies.

In terms of efficiency and ease of implementation, Random Forest demonstrates advantages over SVM. Although SVM can detect more outliers, this algorithm requires precise kernel and parameter selection to achieve optimal performance. This makes SVM implementation more complex. Therefore, in the context of the developed system, Random Forest is considered more efficient and easier to use, especially for fast and stable outlier detection in complex data. Therefore, Random Forest is recommended as a practical and effective outlier detection algorithm for application in customer purchasing behavior data analysis.

References

[1] C. Perdana, U. A. Rosid, and B. A. Okto, “Visualisasi Data Aset Tidak Bergerak Menggunakan Looker Studio Pada PT XYZ,” *J. Inform.*, vol. 3, no. 1, pp. 37–44, 2024, doi: 10.57094/ji.v3i1.1607.

[2] Suhada and M. I. P. Nasution, “Implementasi Teknologi Big Data Dalam Bisnis Untuk Meningkatkan Kepuasan Pelanggan,” *Kohesi J. Multidisiplin Saintek*, vol. 3, no. 7, pp. 23–29, 2024.

[3] D. V. N. Hasibuan and M. I. P. Nasution, “Penerapan Big Data dalam Pemasaran Digital: Studi Kasus pada Industri E-commerce di Indonesia,” *J. Ilm. Nusant. (JINU)*, vol. 1, no. 4, pp. 776–783, 2024, [Online]. Available: https://doi.org/10.61722/jinu.v1i4.1913.

[4] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, “Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya,” *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 3, pp. 307–316, 2023, doi: 10.11594/jesi.02.03.07.

[5] F. Daniel, “Mengatasi Pencilan Pada Pemodelan Regresi Linear Berganda Dengan Metode Regresi Robust Penaksir Lms,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 13, no. 3, pp. 145–156, 2019, doi: 10.30598/barekengvol13iss3pp145-156ar884.

[6] A. Nugroho, M. A. Soeleman, R. A. Pramunendar, A. Affandy, and A. Nurhindarto, “Peningkatan Performa Ensemble Learning pada Segmentasi Semantik Gambar dengan Teknik Oversampling untuk Class Imbalance,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 899–908, 2023, doi: 10.25126/jtiik.20241046831.

[7] Andi, Thamrin, A. Susanto, E. Wijaya, and D. Djohan, “Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease,” *J. Mantik*, vol. 7, no. 2, pp. 2685–4236, 2023, doi: 10.35335/mantik.v7i2.3981.

[8] S. Rahayu and Y. Yamasari, “Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM),” *J. Informatics Comput. Sci.*, vol. 05, no. 03, pp. 440–446, 2024.

[9] A. Wardhana, *Consumer Behavior in The Digital Area 4.0*. Purbalingga: CV. Eureka Media Aksara, 2024.

[10] K. Hayati, A. Tambunan, R. A. Sitorus, and E. S. Sitanggang, “Pengaruh Current Ratio, Inventory Turnover, Total Asset Turnover, dan Debt to Equity Ratio Terhadap Return on Asset Pada Perusahaan Manufaktur Yang Terdaftar di BEI Tahun 2016-2019,” *ASSETS*, vol. 11, no. 2, pp. 221–236, 2021.

[11] F. Diba, M. S. Lydia, and P. Sihombing, “Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi,” *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2152–2160, 2023, doi: 10.33022/ijcs.v12i4.3329.

[12] R. A. Rizal, I. S. Girsang, and S. A. Prasetiyo, “Klasifikasi Wajah Menggunakan Support Vector Machine (SVM),” *REMIK (Riset dan E-Jurnal Manaj. Inform. Komputer)*, vol. 3, no. 2, p. 1, 2019, doi: 10.33395/remik.v3i2.10080.

[13] U. Hanifah, F. Munawarah, and J. Hendra, “Efektivitas Penerapan Sistem Pembayaran Quick Response Code Indonesia Standard (QRIS) pada Bisnis Laundry di Era Modern,” *J. Ekon. dan Kewirausahaan West Sci.*, vol. 2, no. 03, pp. 385–392, 2024, doi: 10.58812/jekws.v2i03.1201.