DOI: 10.14421/ijid.2021.2384 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 10, No. 1, 2021, Pp. 47-52

A Machine Learning Framework for Improving Classification Performance on Credit Approval

Pulung Hendro Prastyo[1], Septian Eko Prasetyo[2], Shindy Arti[3]
Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia
[1]pulung.hendro@mail.ugm.ac.id, [2]septianepe@mail.ugm.ac.id, [3]arti.shindy@mail.ugm.ac.id

Article History
Received Mar 21st, 2021
Revised Jun 27th, 2021
Accepted Jun 28th, 2021
Published Jun, 2021

***Abstract*— Credit scoring is a model commonly used in the decision-making process to refuse or accept loan requests. The credit score model depends on the type of loan or credit and is complemented by various credit factors. At present, there is no accurate model for determining which creditors are eligible for loans. Therefore, an accurate and automatic model is needed to make it easier for banks to determine appropriate creditors. To address the problem, we propose a new approach using the combination of a machine learning algorithm (Naïve Bayes), Information Gain (IG), and discretization in classifying creditors. This research work employed an experimental method using the Weka application. Australian Credit Approval data was used as a dataset, which contains 690 instances of data. In this study, Information Gain is employed as a feature selection to select relevant features so that the Naïve Bayes algorithm can work optimally. The confusion matrix is used as an evaluator and 10-fold cross-validation as a validator. Based on experimental results, our proposed method could improve the classification performance, which reached the highest performance in average accuracy, precision, recall, and f-measure with the value of 86.29%, 86.33%, 86.29%, 86.30%, and 91.52%, respectively. Besides, the proposed method also obtains 91.52% of the ROC area. It indicates that our proposed method can be classified as an excellent classification.**

Keywords*—Loan; Decision-making; Data Mining; Information Gain; Naïve Bayes

1 INTRODUCTION

An evaluation tool that is usually used in the decision-making process is credit scoring model that aims to refuse or accept loan requests [1]. The model is intended to estimate the likelihood of failed customer payment that is evaluated using a credit score. The credit score model depends on the type of loan or credit and complemented by various credit factors, which are entities of actual measurement.

Credit institutions, especially banks, developed the credit scoring model to improve their credit evaluation process and determine creditors' creditworthiness and determine credit risk. Bank is still having difficulty in determining eligible creditors to be given loans or not accurately. Therefore, we need an accurate and automatic credit scoring model to make it easier for banks to determine which creditors are eligible for loans. To address the problem, machine learning algorithm can be used in the data mining process [2].

One of the machine learning algorithms which are convenient, fast in computational time, and requires less data is the Naïve Bayes algorithm [3]–[6]. In the previous study [7], Eweoya et al. employed Naïve Bayes (NB) as a classifier in fraud loan prediction. NB obtained 78% of accuracy and 73.5% of ROC Area. In this study [8], Vimala and Sharmili used Support Vector Machine (SVM) and NB to predict loan risk. NB yielded 77% of accuracy, whereas SVM and SVM+NB obtained 79%. Unfortunately, those results indicate that NB still has low accuracy and needs improvement. Simultaneously, in this study [3][9], the Naïve Bayes algorithm's accuracy is not good enough compared to other machine learning algorithms. Besides, in this study [4][10], the accuracy of Naïve Bayes is also still less than 80%; It is because, in the data mining process, all features are included without selecting relevant features. Therefore, good preprocessing and feature selection is needed to select relevant features, so the classifier's performance can be improved [11]–[14].

Information Gain (IG) is one of the most commonly used filter-based feature selection methods in machine learning. IG is a feature selection method to select relevant features and reduce noisy features [14]–[16]. Therefore, in this study, Information Gain (IG) is employed as a feature selection to improve the performance of the Naïve Bayes algorithm. Combining the Naïve Bayes algorithm and Information Gain is expected to classify creditors precisely and has good accuracy. Discretization is also employed in the preprocessing stage to get better data. Thus, NB can classify data easier.

This study is organized as follows: Section II presents the proposed method. Section III explains the results and analysis. Section IV concludes this research work and provides future works.

2 METHOD

This research work employed an experimental method using the Weka application. The stages of this study can be seen in Figure 1.

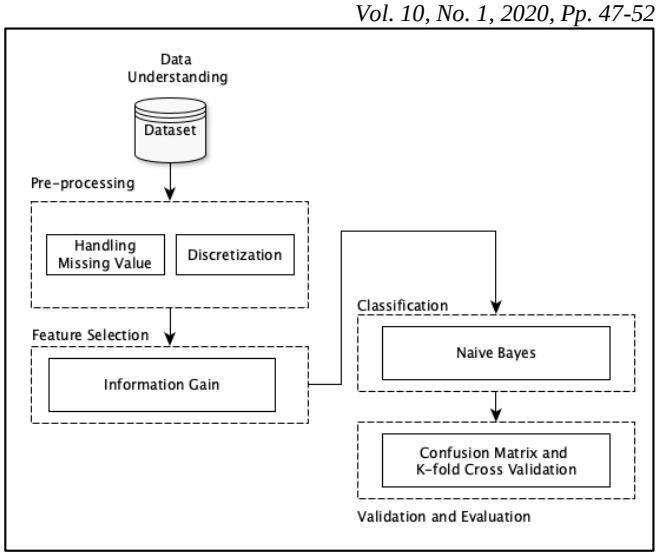

Figure 1. Our proposed method

2.1 Data Understanding

In this study, the Australian Credit Approval data at the UCI Machine Learning Repository is used as a dataset containing 690 instances of data regarding credit card application in Quinlan. The data have 15 features and one target class. All feature names and values were converted to symbols to protect the confidentiality of data.

Table 1. Features of Australian credit approval

Feature	Type	Value
A1	Nominal	a,b
A2	Numeric	13,75 - 80,25
A3	Numeric	0 - 28
A4	Nominal	l,u,y
A5	Nominal	g,p,gg
A6	Nominal	w,q,m,r,cc,k,c,d,x,i,e,aa,ff,j
A7	Nominal	v,h,bb,ff,j,z,o,dd,n
A8	Numeric	0 – 28,5
A9	Nominal	t,f
A10	Nominal	t,f
A11	Numeric	0 - 67
A12	Nominal	t,f
A13	Nominal	g,s,p
A14	Numeric	0 - 2000
A15	Numeric	0 - 100000
Class	Nominal	+,-

Data understanding is essential in order to determine the appropriate pre-processing. Based on the data in Table 1, the nominal type features are 9 data, and the features of the numerical type are 6 data. Data distribution classes (labels) on the Australian Credit Approval can be seen in Table 2.

Table 2. The class of Australian credit approval

Class	Frequency
+	307 (44,5%)
-	383 (55,5%)

Table 2 describes the data class, in which the data can be categorized as balanced data. Positive class data (+) has an amount that is not very different from the negative class data (-). At the same time, the data that consist of the missing value can be seen in Table 3. Missing data can be solved by filling in mode values for nominal data and mean values for

numerical data. The detailed numerical feature data can be seen in Table 4.

Table 3. The missing values of Australian credit approval

Feature	Missing Value
A1	12 (2%)
A2	12 (2%)
A4	6 (1%)
A5	6 (1%)
A6	9 (1%)
A7	9 (1%)
A14	13 (2%)

Table 4. The numerical feature of Australian credit approval

Feature	Minimum Value	Maximum Value	Mean	Standard Deviation
A2	13.75	80.25	31.568	11.958
A3	0	28	4.759	4.978
A8	0	28.5	2.223	3.347
A11	0	67	2.4	4.863
A14	0	2000	184.015	173.807
A15	0	100000	1017.386	5210.103

2.2 Data Preprocessing

In this study, we use two preprocessing stages, namely:

- Handling missing value: to solve the missing value problem. The missing data on the nominal data were replaced by mode data, and mean data replaced the missing data on the numerical data.
- Discretization: In this study, we used the Naïve Bayes algorithm so that the numerical features are changed to nominal. In this paper, we used binning method to change numerical values from features in Table 6 into nominal value divided into 3 groups or bins, such as bin 1, bin 2 and bin 3. Bin 1 until bin 3 are labels. To classify numerical values from A2, A3, A8, A11, A14, and A15 features into bin 1, bin 2, or bin 3, we use formula in Table 5 where x is numerical value.

Table 5. Binning Method

Name	Formula
Bin 1	$x < (Mean - Std.Dev)$
Bin 2	$(Mean - Std.Dev) \leq x \leq (Mean + Std.Dev)$
Bin 3	$x > (Mean + Std.Dev)$

Table 6. Discretized Numerical Features

Features	Labels
A2 A3 A8 A11 A14 A15	Bin 1, Bin 2, and Bin 3

2.3 Information Gain

In this research work, Information Gain (IG) was employed as a feature selection because it can reduce noisy features [15], [16]. IG can detect relevant features for classifiers. To get the best features, the entropy value must be calculated first. Entropy is a measure of class uncertainty using the likelihood of a particular feature. Equation 1 is a formula for calculating entropy. After computing the entropy value, IG can be calculated using Equation 2.

$$Entropy\ (S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (1)$$

Where p_i is the number of samples for class i. Furthermore, c is the number of values in the class classification.

$$Gain\ (S, A) = Entropy(S) - \sum_{values(A)} \frac{|s_v|}{|s|} Entropy(s_v) \quad (2)$$

Where *v* denotes a possible value for feature *A*, *A* is a feature, *Values (A)* is a set of possible values for *A*, |*Sv*| is the number of samples for the value *v*, |*S*| is the totality of all data samples, and *Entropy(Sv)* is the *entropy* for samples that have a value of *v*.

Features which meet the weighting criteria will be employed in the machine learning algorithm. The feature selection by IG is conducted in three stages, namely:

a. Compute IG value for each feature in the used dataset.
b. Determine the required limit (threshold). In this study, top 5 features, 10 features, and threshold (T) ≥ 0.02 were used to examine the best feature combination.
c. The dataset is enhanced by decreasing features used. The results of IG can be seen in Table 7.

Table 7. The Results of Information Gain (IG)

Feature	Value of IG
A1	0.000603
A2	0.019403
A3	0.017232
A4	**0.029603**
A5	0.029603
A6	**0.109160**
A7	**0.050189**
A8	0.017004
A9	**0.425709**
A10	**0.156286**
A11	0.005097
A12	0.000721
A13	0.010036
A14	0.003302
A15	0.005097

2.4 Naïve Bayes

In this study, Naïve Bayes (NB) algorithm is applied to classify creditors. NB is a statistical classification that can be used to predict the probability of membership of a label (class) [17]. Bayes' theorem has a general form like Equation 3.

$$P(h_j|x) = \frac{p(x|h_j)p(h_j)}{p(x)} \quad (3)$$

Where x denotes data with unknown classes, h_j is hypothesis data x, which is a specific class, $P(h_j|x)$ is hypothesis probability h_j based on condition x (posterior probability), $P(h_j)$ is probability h_j hypothesis (prior probability), $p(x|h_j)$ is x probability based on conditions of the hypothesis h_j (likelihood), $p(x)$ is x probability. According to the Bayes formula in Equation 3, we obtain the

Naïve Bayes formula for the feature x_i, which has more than one number n, as in Equation 4 [18].

$$P(h_j|x_1, x_2, .., x_n) = \frac{\prod_{i=1}^{n} p(x_i|h_j)\, P(h_j)}{p(x_1, x_2, .., x_n)} \quad (4)$$

Where Π denotes a multiplication operation. **This formula will be used in the Naïve Bayes algorithm**. We need to remember that the denominator $p(x_1, x_2, .., x_n)$ or the probability of evidence can be eliminated because it only functions as a constant of the same value between classes in its posterior probability. In addition, a class prediction is a class that provides the highest posterior probability values in which the equation can be seen in Equation 5.

$$class\ prediction = \underset{h_j}{\operatorname{argmax}}\ P(h_j) \prod_{i=1}^{n} p(x_i|h_j) \quad (5)$$

2.5 Experimental Setup and Evaluation

This research work is an experimental study as we understand that data mining is an experimental science, in which each algorithm has different results on different data. This research used software and hardware as tools that can be seen in Table 8.

Table 8. Software and hardware

No	Software	Hardware
1	Operation System: OSX High Sierra	CPU: Intel Core i5
2	Weka 3.8.4	RAM: 16 GB
3	Hard Drive	SSD: 240 GB

In this study, we compared a baseline algorithm (Naïve Bayes) with our proposed method. In the proposed method, we employed preprocessing stages, such as handling missing values and discretization. Next, the IG values on the features ranked in the top 5, 10, and the threshold value ≥ 0.02 were employed. Those features were then used in the classification stage. After that, we evaluated the algorithms using accuracy, precision, recall, F-measure, and ROC Area (AUC) [3][19].

The ROC Area illustrates the relationship between the observed class and the predicted class. The accuracy of the ROC classification is done by computing the area under the ROC curve. The area below the curve is called Area Under Curve (AUC). ROC maps two parameters: True Positive Rate and False Positive Rate. AUC delivers an aggregate measure of performance across all possible classification limits. The greater the AUC's value, the better the algorithms [20]. The accuracy criteria for diagnostic tests using AUC are described in Table 9.

Table 9. Criteria of AUC

Value of AUC	Interpretation
90-100	Excellent Classification
80-90	Good Classification
70-80	Fair Classification
60-70	Poor Classification
50-60	Failure

Finally, our proposed method was validated using 10-fold cross-validation. K-fold cross-validation splits training and testing data iteratively as many as *k* values to test the entire data. The experimental scheme in this study can be seen in Figure 2.

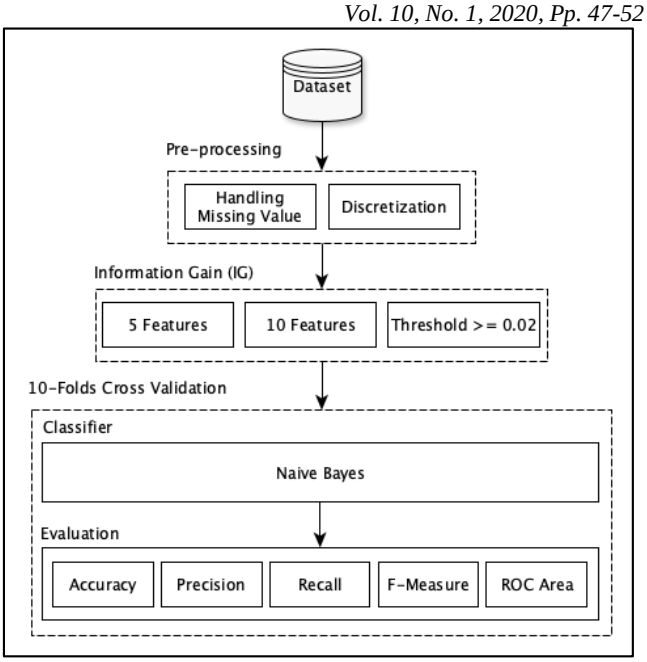

Figure 2. Experimental scheme of this study

3 RESULT AND DISCUSSION

Based on Table 10, it can be seen that the baseline algorithm (Naïve Bayes) obtains the average accuracy, precision, recall, f-measure, and ROC Area with the value of 76.91%, 79.04%, 76.92%, 75.91%, 89.17%, respectively. It indicates that the performance of the Naïve Bayes algorithm is still low. It is similar to the previous studies [7], [8]. In this research work, discretization was employed in the preprocessing stage to make data better. In Table 10, discretization proves to have improved the performance of the Naïve Bayes algorithm significantly. The average accuracy after applying discretization increases by 8.01%, from 77.77% to 85.78%. It indicates that the Naïve Bayes algorithm is very suitable to use nominal data even though it can use numerical data by utilizing the Gaussian function (calculating the mean and variance).

At the same time, the combination of discretization + IG with the five best features can also improve the Naïve Bayes algorithm's performance significantly. The used features are A9, A10, A6, A7, and A4. Based on the findings in this research, the average of confusion matrix value increased by 0.51% from 85.78% to 86.29% for accuracy and recall, precision increased by 0.53% from 85.80% to 86.33%, and f-measure increased by 0.52% from 85.78% to 86.30%. While the ROC area value was 91.52%, which was classified as an excellent classification. This scheme (Discretization + 5 features (IG) + NB) is the best-proposed method visualized in Figure 3.

Table 10. The Comparison of Baseline Algorithm with Our Proposed Method

Algorithm	Model Evaluation				
	Accuracy	Precision	Recall	F-Measure	ROC Area (AUC)
Baseline (NB) [7], [8]	76.91	79.04	76.92	75.91	89.17
Discretization + NB	85.78	85.80	85.78	85.78	91.55
Discretization + 5 Features (IG) + NB	**86.29**	**86.33**	**86.29**	**86.30**	91.52
Discretization + 10 Features (IG) + NB	85.61	85.64	85.63	85.63	**91.63**
Discretization + T ≥ 0.02 Features (IG) + NB	85.55	85.56	85.55	85.55	91.47

In contrast to the use of five features, when using ten features, there was a decrease in the Naïve Bayes algorithm's performance compared to discretization +NB. The features used were A9, A10, A6, A7, A4, A5, A2, A3, A8, and A13. Based on Table 10, the average confusion matrix value decreased by 0.17% for accuracy, and then precision decreased by 0.16%, recall and f-measure decreased by 0.15%. Interestingly, the ROC area increased by 0.08% from 91.55 to 91.63%. The performance of the Naïve Bayes algorithm declined because the noisy features were included in the classification process.

Similar to using ten features, the experiment using threshold ≥ 0.02 also decreased the Naïve Bayes algorithm's performance compared to discretization+NB. However, the decrease was not significant. The features used were six features, namely A9, A10, A6, A7, A4, and A5. Based on Table 10, the average decrease in the model evaluation was 0.23% for accuracy, recall, and f-measure. Then the precision decreased by 0.24%. At the same time, the ROC area decreased by 0.08%. It proves that the selection of appropriate features affects the Naïve Bayes algorithm's performance.

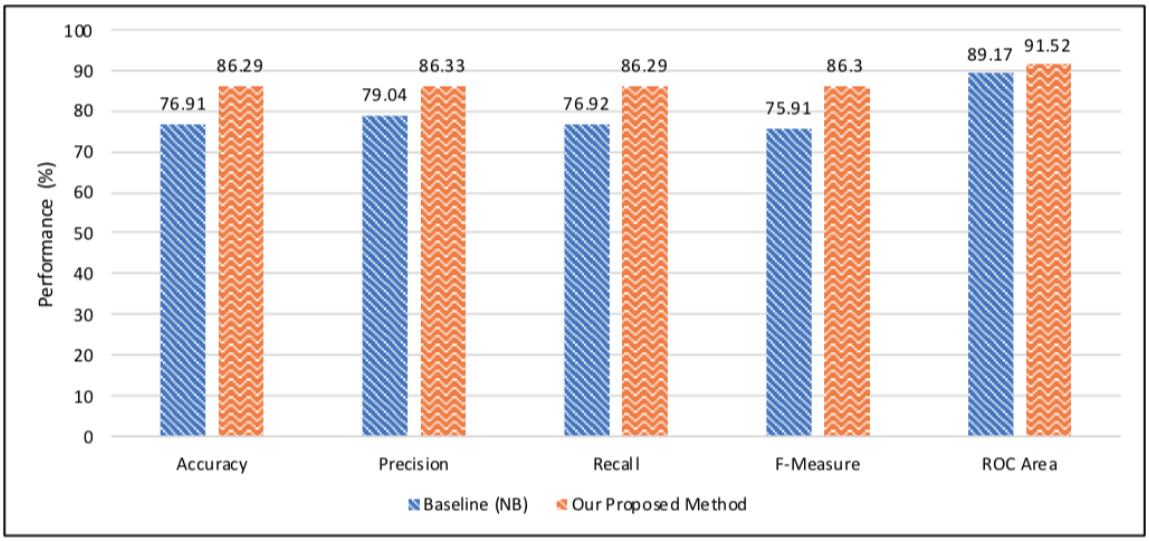

Figure 3. The comparison of baseline algorithm with our best proposed method

4 CONCLUSION

Based on the results of this experiment, our proposed method can improve the classification performance on credit approval, as this method has achieved the best performance with an average accuracy of 86.29%, precision 86.33%, recall 86.29%, f-measure 86.30%. Besides, our proposed method obtained 91.52% of the ROC area. It indicates that our proposed method is classified as an excellent classification. The use of Information Gain is quite sensitive in selecting the best features. It was proven when the ten best features and the threshold ≥ 0.02 (6 features) were used. The algorithm's performance decreased, although it was not too significant. Hence, the selection of the best features using Information Gain should be tested first by using the percentage of the best features (25%, 50%, 75%, and 100%) or using a threshold so that the performance of each feature used can be understood to make it easier to select the suitable features.

For future work, researchers can investigate other preprocessing, feature selection, machine learning algorithms to obtain different results.

ACKNOWLEDGMENT

This work is supported by the Indonesia Endowment Fund for Education, Lembaga Pengelola Dana Pendidikan (LPDP).

REFERENCES

[1] O. J. Okesola, K. O. Okokpujie, A. A. Adewale, S. N. John, and O. Omoruyi, "An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach," in *Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017*, 2017, pp. 228–233, doi: 10.1109/CSCI.2017.36.

[2] Y. Abakarim, M. Lahby, and A. Attioui, "Towards An Efficient Real-time Approach to Loan Credit Approval Using Deep Learning," in *9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018 - Proceedings*, 2018, pp. 306–313, doi: 10.1109/ISIVC.2018.8709173.

[3] P. H. Prastyo, I. G. Paramartha, M. S. M. Pakpahan, and I.

Ardiyanto, "Predicting Breast Cancer : A Comparative Analysis of Machine Learning Algorithms," in *Proceedings International Conference on Science and Engineering*, 2020, pp. 455–459.

[4] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of Naïve Bayes Classification Method for Predicting Purchase," in *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, 2018, pp. 5–9, doi: 10.1109/CITSM.2018.8674324.

[5] F. Burdi, A. H. Setianingrum, and N. Hakiem, "Application of the naive bayes method to a decision support system to provide discounts (Case study: PT. Bina Usaha Teknik)," in *Proceedings - 6th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2016*, 2016, pp. 281–285, doi: 10.1109/ICT4M.2016.57.

[6] A. Tripathi, S. Yadav, and R. Rajan, "Naive Bayes Classification Model for the Student Performance Prediction," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2019*, 2019, pp. 1548–1553, doi: 10.1109/ICICICT46008.2019.8993237.

[7] I. O. Eweoya, A. A. Adebiyi, A. A. Azeta, F. Chidozie, F. O. Agono, and B. Guembe, "A Naive Bayes approach to fraud prediction in loan default," *J. Phys. Conf. Ser.*, vol. 1299, no. 1, p. 4, 2019, doi: 10.1088/1742-6596/1299/1/012038.

[8] S. Vimala and K. C. Sharmili, "Prediction of Loan Risk using Naive Bayes and Support Vector Machine," in *International Conference on Advancements in Computing Technologies*, 2018, pp. 110–113.

[9] R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," in *1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering, ICATIECE 2019*, 2019, pp. 41–45, doi: 10.1109/ICATIECE45860.2019.9063792.

[10] U. Pujianto, E. N. Azizah, and A. S. Damayanti, "Naive Bayes using to predict students' academic performance at faculty of literature," in *5th International Conference on Electrical, Electronics and Information Engineering: Smart Innovations for Bridging Future Technologies, ICEEIE 2017*, 2017, pp. 163–169, doi: 10.1109/ICEEIE.2017.8328782.

[11] J. Ding and L. Fu, "A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search," *J. Intell. Comput.*, vol. 9, no. 3, pp. 93–101, 2018, doi: 10.6025/jic/2018/9/3/93-101.

[12] D. Zeng, J. Peng, S. Fong, Y. Qiu, and R. Wong, "Medical data mining in sentiment analysis based on optimized swarm search feature selection," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 4, pp. 1087–1100, 2018, doi: 10.1007/s13246-018-0674-3.

[13] N. Gopika and A. E. A. Meena Kowshalaya, "Correlation Based Feature Selection Algorithm for Machine Learning," in *Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018*, 2018, pp. 692–695, doi: 10.1109/CESYS.2018.8723980.

[14] M. A. Thanoon, M. J. M. Zedan, and A. N. Hameed, "Feature Selection Based on Wrapper and Information Gain," in *NICST 2019 - 1st Al-Noor International Conference for Science and Technology*, 2019, pp. 32–37, doi: 10.1109/NICST49484.2019.9043805.

[15] S. Widya Sihwi, I. Prasetya Jati, and R. Anggrainingsih, "Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier," in *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*, 2018, pp. 190–195, doi: 10.1109/ISEMANTIC.2018.8549757.

[16] Mihuandayani, E. Utami, and E. T. Luthfi, "Text mining based on tax comments as big data analysis using SVM and feature selection," in *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, 2018, pp. 537–542, doi: 10.1109/ICOIACT.2018.8350743.

[17] P. Chauhan and A. Swami, "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach," *2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018*, pp. 1–8, 2018, doi: 10.1109/ICCCNT.2018.8493927.

[18] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics*, 2nd ed. Penebar Media Pustaka, 2018.

[19] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," in *Procedia Computer Science*, 2018, vol. 127, pp. 511–520, doi: 10.1016/j.procs.2018.01.150.

[20] I. Kurniawati and H. F. Pardede, "Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis," *2018 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2018 - Proc.*, pp. 1–5, 2019, doi: 10.1109/ICITSI.2018.8695953.