IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 4, December 2024, pp. 4364~4371
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i4.pp4364-4371

Predicting levels of legal case difficulties using machine learning

Ilmiyati Sari, Rifki Kosasih, Dina Indarti
Department of Informatics, Faculty of Industrial Technology, Gunadarma University, Depok, Indonesia

Article Info

Article history:

Received Jan 3, 2024
Revised Apr 1, 2024
Accepted Apr 17, 2024

Keywords:

Lawyers
Legal case difficulties
Machine learning
Random forest
Support vector machine

ABSTRACT

Lawyers play a crucial role in the courtroom, assisting clients in their defense. Because of their lack of legal expertise, a person or organization facing legal issues requires professional aid. However, we need to know how much money will be spent on paying lawyers. The level of complexity in a case can be used to determine lawyer costs. Therefore, in this research, we propose employing machine learning methodologies, i.e., random forest classifiers and support vector machines (SVM), to determine the level of legal case difficulties. The novelty of this research is applying a machine learning approach in predicting the level of difficulty of legal cases. The data utilized consists of 990 records, which are divided into training and testing data in a 90:10 ratio. The term frequency-inverse document frequency (TF-IDF) approach was then utilized to perform text preprocessing. The text-preprocessing findings are utilized as input in the classification process. According to the research findings, an accuracy value of 85%, a value of weighted average precision is 88%, and a value of weighted average recall is 85%, for support vector machine. Using random forest, we achieve an accuracy value of 89%, a value of weighted average precision is 85.6%, and a value of weighted average recall is 80%.

Corresponding Author:

Ilmiyati Sari
Department of Informatics, Faculty of Industrial Technology, Gunadarma University
JRJM+974, Pondok Cina, Beji, Depok City, West Java, Indonesia
Email: ilmiyati@staff.gunadarma.ac.id

1. INTRODUCTION

When confronted with a legal situation, most people have to first consult with a lawyer. A person or corporation that is experiencing legal problems needs professional assistance in the legal field because of their limited knowledge of criminal cases. In some legal cases, other legal problems often arise. For example, someone who reports because they feel like a victim may be punished because they could be reported again. Therefore, the services of a lawyer who understands the law well are needed so that these problems can be resolved.

Everything legal can be easily explained and understood with the help of a lawyer. Hiring a lawyer can give numerous benefits, including preventing clients from getting into larger troubles and anticipating the worst-case scenarios that may arise. As a result, engaging a skilled lawyer can assist the client make a better decision and avoid a harsher trial verdict. However, the most important consideration when hiring a lawyer is the cost. The amount of legal fees is typically determined by the difficulty of the case at hand. To determine the level of difficulty of the case being faced, lawyers usually ask questions such as what type of case is being faced, what evidence has been obtained, and the estimated length of time the client will receive the sentence. To help speed up determining the level of difficulty of a case, in this research we propose a machine learning algorithm, i.e., a support vector machine (SVM) and a random forest classifier. Machine learning is an algorithm used to perform learning based on training data. One use of machine learning is for classification, i.e., SVM and random forest.

Several previous studies have made predictions in the legal field, i.e., Şulea *et al.* [1] predicted case decisions at the French supreme court using data from French high court legal decisions using the SVM method with a linear kernel. In his research, f1 score of 75.9% was obtained [1]. Bhilare *et al*. [2] employed descriptive data from legal cases to estimate the result of court decisions using the naive bayes and SVM. In the result, SVM using a linear kernel performed 78% better than naive Bayes [2]. Support vector regression with a linear kernel is also used in research from [3] to identify cases at the European court. The data utilized comprises cases relevant to the convention's paragraphs 3, 6, and 8. According to the research results, an accuracy rate of 79% was achieved [3]. Furthermore, Katz *et al.* [4] employs the random forest to estimate the behavior of the US Supreme Court in a general environment. The case-level and vote-justice data are from the database of the US Supreme Court. According to the research results, at the case level, the value of accuracy is 70.2%, while at the sound fairness level, the value of accuracy is 71.9% [4].

Predictions of court cases in Indonesia were predicted using term frequency-inverse document frequency (TF-IDF) and the k-nearest neighbors (KNN) algorithm in a study [5]. The data utilized is 100 data points from Bandung district court sentence decisions. According to his study, there are four categories of legal cases: small criminal cases, medium criminal cases, huge criminal cases, and corruption cases. An accuracy rate in his research is 86.6667%, a precision value is 85.2862%, and a recall value is 86.6667%. Furthermore, Sari *et al.* [6] employs machine learning for predicting sentence length. The data used consists of 100 sentences of decision data. In his research, cases were grouped into three classes, namely class 1 for imprisonment of less than 5 years, class 2 for imprisonment of 5-10 years, and class 3 for imprisonment of more than 10 years. The longer the prison period they receive, the more complicated the legal cases they face become. The approaches employed include naive Bayes and KNN. The KNN has an accuracy rate is 84%, a precision is 90%, and a recall is 84%. The naive Bayes classifier has an accuracy rate is 80%, a precision is 72.16844%, and a recall is 80%.

Based on previous research, SVM can be used to predict or identify legal cases [1], [2]. Meanwhile, alternative machine learning algorithms, i.e., naive Bayes and KNN, have been used in a study [6] to estimate the length of punishment using 100 data. This study continues the research of [6], which recommended employing another machine learning algorithms, i.e., SVM and random forest, to predict the level of difficulty of a legal case. The data used in this research was expanded to 990 legal decision data containing evidence used in the case.

2. METHOD

To predict the level of difficulty of a legal case, numerous stages are required, as shown in Figure 1. We collected 990 data from the Bandung district court. This data represents case decisions at the Bandung District Court. The decision contains the attributes of the evidence and the length of the defendant's punishment. These attributes will be used as features to predict the difficulty level of a case.

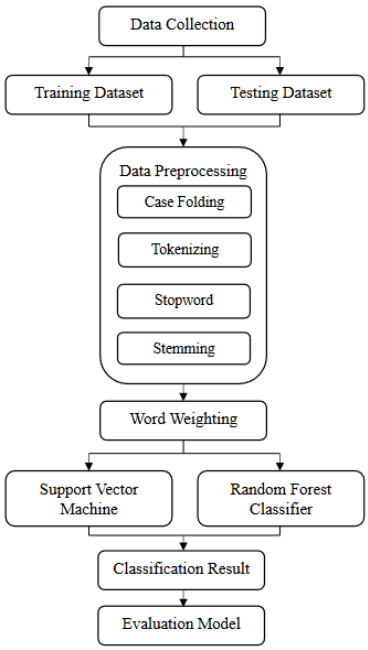

Figure 1. The flowchart of the AI-based models and experimental methods applied

The data is divided into 4 categories based on the level of case difficulty. We divide the data into a training and a testing dataset with a ratio of 90:10. The next stage is to carry out preprocessing on the sentence decision data so that the data can be used.

2.1. Data preprocessing

Data preprocessing is used to remove noise from the data. In this study, we use four steps, i.e., case folding, tokenizing, stopping words, and stemming. The case folding stage involves changing capital letters to lowercase. The next stage is the tokenizing stage, which involves breaking down sentences into single words [7]. The next stage is to do filtering by performing stopword removal, which means that every word generated by tokenizing is reviewed, and any prepositions that have nothing to do with analysis are eliminated [8]. Stemming is the last stage of preprocessing. In this step, words are changed to basic words [9]. In the next stage, we perform weighting of words.

2.2. Weighting of word

Weighting of words is the process of assigning weights to each word by using the TF-IDF approach, as in (1) and (2), so that it may be utilized as input in the classification. The phases of word weighting using TF-IDF are as follows [10]. The initial step in word weighting is to compute the term frequencies (TF) for each word. Sentences that have been broken down into words will be scored. Each word will be assigned a value of 1. The following step is to determine the document frequencies (DF) for each word by summing the TF values for each word. The third stage is to compute the inverse document frequencies (IDF) as in (1).

$$IDF(W) = \log\left(\frac{N}{DF(W)}\right) \quad (1)$$

After that, we assign a weight to each word. We multiply the TF value by the IDF as in (2).

$$W_{ij} = tf_{ij} \times \log\frac{D}{df_j} \quad (2)$$

The word weighting results will be utilized as input for the classification process, which will employ the SVM and random forest classifier.

2.3. Support vector machine

In SVM, we find the hyperplane in the feature space that best divides the data into distinct classes [11], as shown in (3).

$$f(x) = w \cdot x + b \quad (3)$$

where $f(x)$ is subspace of set of hyperplanes, x is space of input, w is weighted vector, and b is bias. A line is called a hyperplane in two dimensions. It refers to subspace in higher dimensions. The margin is the distance between the hyperplane and the closest data point from each class, and SVM seeks the hyperplane that maximizes this [12]–[15]. As seen in Figure 2, the closest data point is called the support vector.

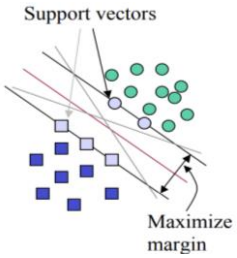

Figure 2. The mechanism of SVM

By utilizing a kernel trick, SVM can handle non-linear decision boundaries [16]–[19]. The features of the input must be mapped in order to construct a linear decision boundary. In SVM, there are three kernel functions, i.e., linear, the radial basis function (RBF) and polynomial kernels [20]–[22]. In practical scenarios

where the data exhibits some overlap or is not entirely separable, SVM can be extended to accommodate a certain degree of misclassification. This extension is known as a "soft margin," and the SVM algorithm strives to find a balance between maximizing the margin and minimizing misclassifications. Once the hyperplane is established, the process of classifying a new data point involves determining on which side of the hyperplane it resides [23], [24]. SVM find extensive applications in diverse fields such as image classification, text categorization, and bioinformatics. Their effectiveness lies in their adeptness at managing intricate relationships within data and exhibiting robust generalization to previously unseen data.

2.4. Random forest classifier

A random forest ensemble learning approach is used for classification. It generates a huge number of decision trees during training, then produces the mean prediction, or the mode of the classes for each tree [25]–[28]. A comprehensive outline of the random forest algorithm is provided:

- Random sampling with replacement using bootstraps: choose a portion of the training data at random using replacement. This implies that certain cases might be included twice in the subset and others might not.
- Feature randomization: for each decision tree node, only a random feature is considered for splitting. This contributes to the introduction of tree diversity.
- Building decision trees: develop a decision tree for each collection of features and data. Recursively, growing the tree involves periodically dividing the data according to the chosen features until a halting condition is satisfied (e.g., minimum samples per leaf and maximum depth attained).
- Voting or averaging: in classification, each tree "votes" for a class, and the class with the most votes become the predicted class. To determine the final regression prediction for a given regression task, the predictions made by each tree are averaged.
- Collective outcome: the combined outcome of each individual tree is the ultimate output. Through the reduction of overfitting and the capture of a more robust pattern in the data, this ensemble technique contributes to the enhancement of the model's resilience and generalization.

The benefits of random forest are high accuracy, robustness to overfitting, and feature importance. The drawbacks of using random forest are complexity and computational cost [29]. Random forests are a widely used approach because they may be used for a wide range of activities and data types. They are widely used in machine learning due to their robustness and ability to manage complex data interactions.

2.5. Evaluation model

We evaluate the model by calculating accuracy (Acc), precision and recall as in (4) to (6) [30]–[32]. Accuracy is one of the metrics used to measure the accuracy of data by calculating the comparison between the correct predictions and the entire data. Precision is the comparison between true positive predictions and the overall positive predicted results, while recall is the comparison between true positive predictions and all true positive data. For multi-class cases, weighting is needed to calculate precision and recall using weighted average precision (WAP) and weighted average recall (WAR) as in (7) and (8).

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (4)$$

$$p_k = \frac{TP_k}{TP_k+FP_k} \times 100\% \quad (5)$$

$$WAP = \frac{\sum_{k=1}^{n} p_k d_k}{\sum_{k=1}^{n} d_k} \quad (6)$$

$$r_k = \frac{TP_k}{TP_k+FN_k} \times 100\% \quad (7)$$

$$WAR = \frac{\sum_{k=1}^{n} r_k d_k}{\sum_{k=1}^{n} d_k} \quad (8)$$

with TP and TN are true positive and true negative, FP and FN are false positive and false negative, p_k is precision in k class, d_k is actual data in the k class, and r_k is recall in k class.

3. RESULTS AND DISCUSSION

In our research, we collect 990 data on sentence decisions at the Bandung District Court. We divided the data into two datasets: training and testing, with a ratio of 90:10. 1 contains examples of the data that has been collected. Table 1 data from sentencing decisions at the bandung district court (4 of 990 dataset).

Table 1. Dataset of sentencing decisions at the bandung district court (4 of 990 dataset)

Decision	Code
Stated that the defendant GREDY PANTERA Als IYANG Bin GUGUN GUNAWAN was proven to have committed the criminal act "Attempt or evil conspiracy to commit a Narcotics crime without any right or against the law to possess, keep in possession of or provide Class I Narcotics which are not plants as intended in paragraph (1) weighing more than 5 (five) grams" 2. The defendant GREDY PANTERA Als IYANG Bin GUGUN GUNAWAN was sentenced to imprisonment for 7 (seven) years and a fine of Rp. 1,000,000,000. 3. State evidence in the form of; 1 box of Sampoerna Mild cigarettes contains two small plastic clip packages of crystal methamphetamine weighing 0.3260 grams. 20 (twenty) small plastic clip packages of Narkotija type crystal methamphetamine weighing 5.2525 grams. 1 unit of Oppo brand cellphone;	3
Stated that the Defendants: 1. Asep Warman bin Mamat, 2. Yadi Setiadi alias Tahu bin Mamat Rahmat, 3. Tatang Supriadi bin Mansyur S. mentioned above, were proven to have committed "violence against people who caused injury" as in the first alternative indictment; Impose prison sentences of 1 (one) year and 4 (four) months respectively; Establish as evidence 1 BMC brand helmet and 1 Double Stick;	1
Stated that the Defendant AGUS SUHERMAN Bin HARUN KOMARUDIN was proven to be Circulating Counterfeit Rupiah ". Sentenced AGUS SUHERMAN Bin HARUN KOMARUDIN to 2 years and fined Rp. 200,000,000, Order the defendant to remain in detention; evidence in the form of: Fake rupiah currency amounting to Rp. 10,000,000- which is still tied with a money rope, Rp. 100,000 for 100 shares;	2
DECIDES Declaring that the prosecution of the Defendant YAYAH WILIYAH Binti ALM M. HARIS is discontinued because the person concerned has died; Order the Registrar to record the dismissal of the prosecution against the defendant YAYAH WILIYAH Binti ALM M. HARIS in the Criminal case register Number: 54/Pid.B/2022/PN. Bdg;	4

To predict the level of difficulty of a legal case, text preprocessing is first carried out. The first step is case folding, which converts all letters to lowercase. Tokenizing involves breaking down sentences into single words. The next step is the stopword stage, basically, each word is reviewed, and if there are words that have nothing to do with analysis, will be eliminated. Finally, during the stemming step, the collection of words processed with stop words is transformed into the form of basic words or affixes are removed. Table 2 displays the outcomes of data preprocessing.

Table 2. Result of data preprocessing (4 of 990 dataset)

Decision	Code
indictment of Gredy Pantera Als Iyang bin Gugun Gunawan, evidence of a criminal act, attempted conspiracy, criminal act of narcotics, Haka, you are against the law, possession, possession of power of attorney for narcotics, class of planting, paragraph 1, weighing more than 5, five grams, 2 convictions, indictment of Gredy Pantera Als Iyang bin Gugun Gunawan, imprisonment for 7 seven year fine Rp. 1,000,000,000 charge order to detain order to detain 3 items of evidence 1 one box of Sampoerna mild cigarettes containing two plastic packets of narcotic clips, type of methamphetamine, heavy 0 3260 grams 20 twenty packages of plastic clips of narcotic narcotics, type of methamphetamine, heavy 5 2525 grams 1 unit Oppo brand cell phone	3
Indictment 1 Asep Warman Bin Mamat 2 Yad Setiadi alias Bin Mamat Rahmat 3 Tatang Supriadi Bin Mansyur S on evidence of criminal acts of violent behavior of people due to injuries Alternative charges one fall criminal charge prison sentence 1 year 4 four months still arrest detained on the road Charge less criminal fall Still accused, still hold evidence in the form of 1 BMC brand helmet, 1 double stick	1
Indicted Agus Suherman bin Harun Komarudin, legal evidence of wrongdoing in the crime of circulating counterfeit rupiah, convicted. Indicted by Agus Suherman bin Harun Komarudin, 2-2 years imprisonment, fine of IDR 200,000,000, of course the fine will be paid in lieu of imprisonment for 2 months, still detained on the road. Insufficient sentence falls. Order to indict detained. remains evidence of fake rupiah currency Rp. 10,000,000 - tie a string tied broken money Rp. 100,000 - 100 pieces	2
Yayah Wiliyah Binti Alm M Haris has died and remains involved in the world, the clerk's order is to record the failure to prosecute Yayah Wiliyah Binti Alm M Haris, criminal case registration number 54 PID B 2022 PN BDG	4

Later, we perform word weighting by using the TF-IDF method, as in (1) and (2). This result is input for the classification process. To perform classification, we employ SVM and random forest classifiers. Tables 3 and 4 present the classification results.

Table 3. Confusion matrix of SVM

		Prediction			
Actual	Class	1	2	3	4
	1	40	1	0	1
	2	8	30	2	1
	3	1	1	13	0
	4	0	0	0	1

In Table 3, in the actual situation, as many as 40 legal cases with a class 1 difficulty level are predicted to enter class 1, 1 case is predicted to enter class 2, and 1 case is predicted to enter class 4. For legal cases with class 2 difficulty level, in actual circumstances, 30 cases are predicted to enter class 2, 8 cases are predicted to enter class 1, 2 cases are predicted to enter class 3 and 1 case is predicted to enter class 4. For legal cases with a difficulty level of class 3, in actual circumstances, 13 cases are predicted to enter class 3, 1 is predicted to enter class 1 and 1 is predicted to enter class 2. Meanwhile, for cases with a class 4 difficulty level, it is predicted to enter class 4 as many as 1 case.

According to Table 4, in the actual situation, 44 legal cases with a class 1 difficulty level are predicted to enter class 1, and 1 case is predicted to enter class 2. In actual circumstances, 30 legal cases with a difficulty level of class 2 are predicted to enter class 2, 4 cases are predicted to enter class 1, 2 cases are predicted to enter class 3, and 1 case is predicted to enter class 4. In actual circumstances, 13 legal cases with a difficulty level of class 3 are anticipated to enter class 3, 1 is predicted to enter class 1, 1 is predicted to enter class 2, and 1 is predicted to enter class 4. Meanwhile, for cases with a class 4 difficulty level, it is predicted to enter class 4 as many as 1 case. The model is then evaluated by computing the accuracy value, weighted average of precision, and weighted average of recall with (4), (6), and (8). Table 5 summarizes the model evaluation results.

Table 4. Confusion matrix of random forest

		Prediction			
	Class	1	2	3	4
Actual	1	44	1	0	0
	2	4	30	2	1
	3	1	1	13	1
	4	0	0	0	1

Table 5. Result of evaluation model

Class	SVM			Random forest			Number of test data
	Acc (%)	Precision (%)	Recall (%)	Acc (%)	Precision (%)	Recall (%)	
1		82	95		90	98	4
2		94	74		94	82	1
3		87	87		87	81	3
4		0	0		0	0	4
	85			89			
Weighted Average		88	85		91	89	

Table 5 summarizes the overall evaluation results. The model, which was constructed with a SVM, has accuracy value of 85%, a WAP of 88%, and a WAR of 85%. Using random forest, we achieve an accuracy value of 89%, a value of WAP is 85.6%, and a value of WAR is 80%. Based on these results, it can be concluded that the random forest method has better results than using the SVM method in predicting the level of difficulty of a legal case.

4. CONCLUSION

One factor in determining attorney fees is the level of difficulty of the legal case being faced. The more difficult the case is to resolve, the more expensive it will be to pay a lawyer. Therefore, it is very important for us to know the level of difficulty of the legal case we are facing. In this research, we propose SVM and random forest classifiers to estimate the level of difficulty of a legal case. The data used consists of 990 data. We divide the data with a ratio of 90:10 for training data and testing dataset. After that, text preprocessing was carried out using the TF-IDF method. The results of the preprocessing text are used as input in the classification process. Based on the research results, an accuracy value of 85%, a value of WAP is 88%, and a value of WAR is 85%. Using random forest, we achieve an accuracy value of 89%, a value of WAP is 85.6%, and a value of WAR is 80%. In the case of this data, it can be concluded that the random forest method is a better model than the SVM for predicting the level of difficulty of a legal case. For further research, other methods can also be used as a comparison, such as a deep learning approach.

ACKNOWLEDGEMENTS

This research was funded by PTUPT 2023 Research Grant, Ministry of Education, Culture, Research and Technology, Indonesia to Gunadarma University No. 073/E5/PG.02.00.PL/2023 in 12 April 2023.

REFERENCES

[1] O.-M. Șulea, M. Zampieri, M. Vela, and J. V. Genabith, "Predicting the law area and decisions of French supreme court cases," *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 716–722, 2017, doi: 10.26615/978-954-452-049-6_092.

[2] P. Bhilare, N. Parab, N. Soni, and B. Thakur, "Predicting outcome of judicial cases and analysis using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 326, pp. 326–330, 2008.

[3] N. Aletras, D. Tsarapatsanis, D. P. -Pietro, and V. Lampos, "Predicting judicial decisions of the European court of human rights: A natural language processing perspective," *PeerJ Computer Science*, vol. 2016, no. 10, pp. 1–19, 2016, doi: 10.7717/peerj-cs.93.

[4] D. M. Katz, M. J. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," *PLoS One*, vol. 12, no. 4, pp. 1–18, 2017, doi: 10.1371/journal.pone.0174698.

[5] I. Sari and R. Kosasih, "Prediction types of legal cases in indonesia using the TF-IDF method and the KNN algorithm," *AIP Conference Proceedings*, vol. 2689, no. 1, 2023, doi: 10.1063/5.0114543.

[6] I. Sari, R. Kosasih, and A. Fahrurozi, "Implementation of machine learning in predicting length of punishment at Bandung Court," in *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 2022, pp. 12–17, doi: 10.1109/ICSINTESA56431.2022.10041693.

[7] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique," *2016 4th international conference on information and communication technology (ICoICT)*, 2016, doi: 10.1109/ICoICT.2016.7571912.

[8] N. Normah, "Naïve Bayes algorithm for sentiment analysis windows phone store application reviews," *SinkrOn*, vol. 3, no. 2, 2019, doi: 10.33395/sinkron.v3i2.242.

[9] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, 2016, doi: 10.21512/comtech.v7i4.3746.

[10] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0192-7.

[11] I. L. Mahargya and G. F. Shidik, "Improvement support vector machine using genetic algorithm in farmers term of trade prediction at Central Java Indonesia," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 6, pp. 2261–2269, 2020, doi: 10.18517/ijaseit.10.6.9400.

[12] Y. B. Salem, M. N. Abdelkrim, and Y. B. Salem, "Texture classification of fabric defects using machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4390–4399, 2020, doi: 10.11591/ijece.v10i4.pp4390-4399.

[13] N. N. Moon *et al.*, "Natural language processing based advanced method of unnecessary video detection," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5411–5419, 2021, doi: 10.11591/ijece.v11i6.pp5411-5419.

[14] I. S. A. -Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, pp. 5497–5506, 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.

[15] L. K. Ramasamy, S. Kadry, Y. Nam, and M. N. Meqdad, "Performance analysis of sentiments in twitter dataset using SVM models," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2275–2284, 2021, doi: 10.11591/ijece.v11i3.pp2275-2284.

[16] I. M. Hayder, G. Abdul, N. Al, and H. A. Younis, "Predicting reaction based on customer's transaction using machine learning approaches," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1086–1096, 2023, doi: 10.11591/ijece.v13i1.pp1086-1096.

[17] Y. K. Zamil, S. A. Ali, and M. A. Naser, "Spam image email filtering using K-NN and SVM," *International Journal of Electrical and Computer Engineering* , vol. 9, no. 1, pp. 245-254, 2019, doi: 10.11591/ijece.v9i1.pp245-254.

[18] G. W. -Wen, Y. Lv, Y. J. Yu, Z. Wang, and S. Y. -Hai, "Fast support vector classifier with generalization-memorization kernel," *Procedia Computer Science*, vol. 214, pp. 55–62, 2022, doi: 10.1016/j.procs.2022.11.148.

[19] A. B. Gumelar, A. Yogatama, D. P. Adi, Frismanda, and I. Sugiarto, "Forward feature selection for toxic speech classification using support vector machine and random forest," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 717–726, 2022, doi: 10.11591/ijai.v11.i2.pp717-726.

[20] B. Abuhaija *et al.*, "A comprehensive study of machine learning for predicting cardiovascular disease using Weka and SPSS tools," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1891–1902, 2023, doi: 10.11591/ijece.v13i2.pp1891-1902.

[21] B. T. Khoa and T. T. Huynh, "Forecasting stock price movement direction by machine learning algorithm," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 6, pp. 6625–6634, 2022, doi: 10.11591/ijece.v12i6.pp6625-6634.

[22] M. D. Salawu *et al.*, "A chi-square-SVM based pedagogical rule extraction method for microarray data analysis," *International Journal of Advances in Applied Sciences*, vol. 9, no. 2, pp. 93–100, 2020, doi: 10.11591/ijaas.v9.i2.pp93-100.

[23] A. Susanto, C. A. Sari, H. Rahmalan, and M. A. S. Doheir, "Support vector machine based discrete wavelet transform for magnetic resonance imaging brain tumor classification," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 21, no. 3, pp. 592–599, 2023, doi: 10.12928/TELKOMNIKA.v21i3.24928.

[24] C. B. Tan, M. H. A. Hijazi, and P. N. E. Nohuddin, "A comparison of different support vector machine kernels for artificial speech detection," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 21, no. 1, pp. 97–103, 2023, doi: 10.12928/TELKOMNIKA.v21i1.24259.

[25] V. Y. Kullarni and P. K. Sinha, "Random forest classifier: a survey and future research directions," *International Journal of Advanced Computer Technology*, vol. 36, no. 1, pp. 1144–1156, 2013.

[26] L. Breiman, "Bagging predictors," in *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/bf00058655.

[27] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 1208–1215, 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.

[28] M. H. Mutar, E. H. Ahmed, M. R. M. ALsemawi, H. O. Hanoosh, and A. H. Abbas, "Ear recognition system using random forest and histograms of oriented gradients techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 181–188, 2022, doi: 10.11591/ijeecs.v27.i1.pp181-188.

[29] A. A. Khaleel, A. A. M. A.-Azzawi, and A. M. Alkhazraji, "Random forest for lung cancer analysis using apache mahout and hadoop based on software defined networking," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1086–1093, 2023, doi: 10.11591/ijeecs.v32.i2.pp1086-1093.

[30] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Computers & Geosciences*, vol. 184, 2024, doi: 10.1016/j.cageo.2024.105531.
[31] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 18, no. 2, pp. 815–821, 2020, doi: 10.12928/TELKOMNIKA.v18i2.14785.
[32] S. O. -Arias, J. S. Piña, R. T. -Soto, L. F. C. -Ossa, R. Guyot, and G. Isaza, "Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements," *Processes*, vol. 8, no. 6, 2020, doi: 10.3390/PR8060638.

BIOGRAPHIES OF AUTHORS

Ilmiyati Sari received her B.S. degree in Mathematics from University of Indonesia in 2009, her M.S. in Mathematics from University of Indonesia in 2012 and her doctoral in Gunadarma University in 2018. She has published extensively in the area of video processing. Her research interests include dynamic models, statistics, image processing, video processing, and machine learning. She can be contacted at email: ilmiyati@staff.gunadarma.ac.id.

Rifki Kosasih received the B.S. degree in Mathematics from University of Indonesia, Indonesia, in 2009, the M.S. degree in Mathematics from University of Indonesia in 2012, and the Ph.D. degree from Gunadarma University, Indonesia, in 2015. He is a lecturer in Department of Informatics (Gunadarma University). His research interests are image processing, object recognition, data science, manifold learning, machine learning, and deep learning. He can be contacted at email: rifki_kosasih@staff.gunadarma.ac.id.

Dina Indarti received her B.S. degree in Mathematics from University of Indonesia in 2008, her M.S. in Mathematics from University of Indonesia in 2011 and her doctoral in Gunadarma University in 2014. She has published extensively in the area of image processing. Her research interests include statistics, image processing, and machine learning. She can be contacted at email: dina_indarti@staff.gunadarma.ac.id.