JSAI : Journal Scientific and Applied Informatics
Vol. 6, No. 3, November 2023, hal. 493-498
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Pengaruh *Tuning Parameter* dan *Cross Validation* Pada Klasifikasi Teks Komplain Bahasa Indonesia Menggunakan Algoritma *Support Vector Machine*

Vina Ayumi[1a], Desi Ramayanti[2b], Handrie Noprisson[1c], Anita Ratnasari[1d], Umniy Salamah[1e]
[1]Fakultas Ilmu Komputer, Universitas Mercu Buana, Jakarta, Indonesia
[2]Fakultas Teknik dan Informatika, Universitas Dian Nusantara, Jakarta, Indonesia
[a]*vina.ayumi@mercubuana.ac.id,* [b]*desi.ramayanti@undira.ac.id,* [c]*handrie.noprisson@mercubuana.ac.id,*
[d]*anita.ratnasari@mercubuana.ac.id,* [e]*umniy.salamah@mercubuana.ac.id*

Article Info

Article history:

Received, 2023-04-16
Revised, 2023-10-18
Accepted, 2023-11-20

Kata Kunci:

Komplain Masyarakat
Tuning Parameter
Cross Validation
Support Vector Machine

ABSTRAK

Klasifikasi teks bertujuan untuk mengelompokkan data teks, misalnya, untuk menemukan beberapa informasi dari dataset teks dari media sosial yang berukuran besar sehingga dapat digunakan oleh pemilik data (*data ow*ner). Klasifikasi teks secara manual memakan waktu dan sulit sehingga beberapa peneliti mencoba untuk melakukan riset klasifikasi teks secara otomatis. Penelitian ini mencoba untuk mengklasifikasikan dataset teks Bahasa Indonesia dengan menggunakan algoritma SVM. Penelitian dilakukan dalam dua tahap, yaitu eksperimen pertama tanpa parameter *cross validation* dan *tuning parameter*, kemudian dilakukan eksperimen kedua dengan parameter *cross validation* dan *tuning parameter*. Eksperimen tanpa parameter *cross validation* dan *tuning parameter* untuk support vector machine (SVM) mendapatkan akurasi 89,47% dengan nilai *precision* dan *recall* masing-masing adalah 0,90 dan 0,89. Eksperimen kedua menggunakan *cross validation* dengan *k-5* dan *k-10* dan *parameter tuning* dengan *C constant* dan *gamma value*. Hasil *cross validation* dengan *k-10* diperoleh akurasi terbaik dengan nilai 96,48% denganwaktu komputasi selama 40,118 detik. Selanjutnya, fungsi kernel pada *parameter tuning* yaitu *sigmoid*, *linear* dan *radial basis function* dianalisis dan didapatkan bahwa fungsi kernel *sigmoid* mencapai akurasi dan waktu komputasi terbaik.

Keywords:

Community Complaint
Tuning Parameter
Cross Validation
Support Vector Machine

ABSTRACT

Text classification aims to group text data, for example, to find some information from a large social media text dataset so that it can be used by the data owner. Manual text classification is time-consuming and difficult, so some researchers try to research text classification automatically. This study attempts to classify Indonesian text datasets using the SVM algorithm. The research was conducted in two stages, namely the first experiment without cross validation parameters and parameter tuning, then the second experiment was carried out with cross validation parameters and parameter tuning. Experiments without cross validation parameters and parameter tuning for support vector machines (SVM) obtained 89.47% accuracy with precision and recall values of 0.90 and 0.89 respectively. The second experiment used cross validation with k-5 and k-10 and tuning parameters with C constant and gamma values. Cross validation results with k-10 obtained the best accuracy with a value of 96.48% with a computation time of 40.118 seconds. Next, kernel functions in tuning parameters namely sigmoid, linear and radial basis functions are analyzed and it is found that sigmoid kernel functions achieve the best accuracy and computational time.

Penulis Korespondensi:

Vina Ayumi,
Fakultas Teknik dan Informatika
Universitas Mercu Buana, Jakarta, Indonesia
Email: vina.ayumi@mercubuana.ac.id

JSAI : Journal Scientific and Applied Informatics
Vol. 6, No. 3, November 2023, hal. 493-498
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

1. PENDAHULUAN

Pattern recognition merupakan salah satu topik penelitian yang dapat diadaptasi untuk pengolahan data citra maupun data teks. Banyak algoritma yang dapat diterapkan untuk *pattern recognition* salah satunya *support vector machine* [1], [2]. Konsistensi penggunaan SVM dalam penelitian ditunjukkan dari penelitian saat ini yang masih menggunakan metode tersebut untuk menyelesaikan permasalahan yang ada di penelitian [3]–[5].

Metode *support vector machine* (SVM) merupakan algoritma *machine learning* yang diusulkan oleh Vladimir Vapnik [6]. SVM memiliki tujuan untuk mencari bidang pemisah data secara optimal (*hyperplane*) dalam sebuah *input space*. Sebuah *hyperplane* yang terbaik didapatkan dengan mengukur *margin* maksimal antara batasan kelas pada ruang input. Cara kerja support vector machine (SVM) bersifat *linear classifier*, namun algoritma ini dapat digunakan untuk *problem linear* dengan memanfaatkan konsep *kernel trick* dengan cara memetakan *input space* ke ruang *high-dimensional* [6]. Selain itu, SVMs memiliki kelebihan yang dapat secara independen belajar dari *dimensional feature space*.

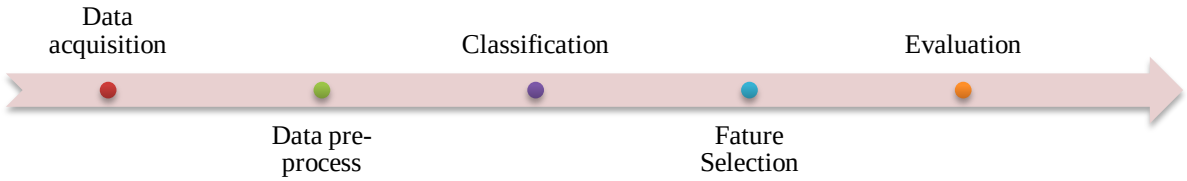
Penelitian oleh [7] menggunakan SVM untuk klasifikasi data *high dimensional* dan seleksi fitur. Pada bidang keilmuan lain, SVM masih digunakan menyelesaikan permasalahan yang ada. Melihat konsistensi penggunakan metode SVM dalam berbagai kasus, penelitian ini mencoba untuk menerapkan *support vector machine* pada proses klasifikasi teks Bahasa Indonesia setelah dataset teks dilakukan pra-proses terlebih dahulu [8]–[10]. Selanjutnya, untuk meningkatkan akurasi SVM bisanya dilakukan *tuning parameter*. Kernel *sigmoid*, *radial basis function*, dan *linier* merupakan parameter di *support vector machines* (SVM) penting untuk meningkatkan akurasi klasifikasi. SVM *tuning parameter* telah ditemukan berkinerja baik dalam hal akurasi klasifikasi [11], [12], [21]–[23], [13]–[20].

Selain itu, *cross validation* penting untuk membantu menentukan nilai parameter optimal untuk fungsi kernel untuk mendapatkan akurasi maksimum dalam tugas klasifikasi menggunakan SVM. Metode *cross validation* juga sangat penting untuk memvalidasi akurasi karena dapat digunakan untuk menghitung kesalahan *cross validation* pada *global minimum* secara efisien, sehingga kemampuan generalisasi menjadi lebih baik dan waktu komputasi yang lebih rendah dibandingkan dengan metode lain. Secara keseluruhan, metode cross *validation* memainkan peran penting dalam mengoptimalkan nilai parameter, memvalidasi kinerja sistem, dan meningkatkan efisiensi model SVM [24]–[35].

Berdasarkan latar belakang diatas, penelitian ini mengusulkan penelitian untuk mengetaui pengaruh *tuning parameter* dan *cross validation* pada klasifikasi teks komplain Bahasa Indonesia menggunakan algoritma *support vector machine*.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan melalui lima tahapan utama, antara lain data acquisition, data pre-process, feature selection, classification dan evaluation seperti pada Gambar 1.



Gambar 1 Tahap Penelitian

Pada tahap akuisisi dilakukan dengan cara mengekstrak data secara otomatis menggunakan script program di R melalui Twitter API dan didapatkan data sebanyak 1170 data tweet yang terdiri dari data komplain dan bukan komplain. Selanjutnya, tahap praproses dilakukan dengan cara *data cleansing* dengan menghapus data yang duplikat (retweet), serta menghapus repetisi kata di dalam tweet, *data labelling* dengan memberikan label untuk data yang tergolong komplain dan bukan komplain, c*ase folding* dengan mengkonversi seluruh teks kedalam bentuk *lower case* dan r*emoving special character* dan s*top stopword*, menghilangkan kata-kata yang dianggap tidak memiliki makna.

Tahap selanjutnya adalah tahap seleksi fitur yaitu memilih fitur yang tepat dan membentuk *feature vector*. Pada penelitian ini kami menggunakan metode *term frequency-inverse document frequency* (TF-IDF) untuk merepresentasikan data. Setelah dikonversi menjadi vektor selanjutnya adalah pemisahan data kedalam data pelatihan dan pengujian. Setelah itu dilakukan tahap klasifikasi. Tahap ini dilakukan dengan melatih SVM classisfier dengan data pelatihan sehingga membentuk model. Setelah dihasilkan model kemudian model diuji dengan data pengujian untuk memprediksi label datanya. Proses klasifikasi dapat asumsikan data yang diolah dapat dipisahkan secara linear (*linearly separable*) sehingga support vector machine (SVM) memisahkan data dengan sesuai dengan *hyperplane* dengan persamaan [6]:

$$x.w + b = 0$$

Dimana

- w adalah vector bobot
- b suatu scalar.

Sebuah *hyperplane* tersebut akan memisahkan data menjadi dua kelas yaitu kelas negative dan positif dengan persamaan [6]:

$$x_i.w + b \geq -1 \text{ untuk kelas negatif } y_i = -1$$
$$x_i.w + b \geq 1 \text{ untuk kelas postif } y_i = 1$$

Sehingga dapat disimpulkan dalam persamaan berikut [6]:

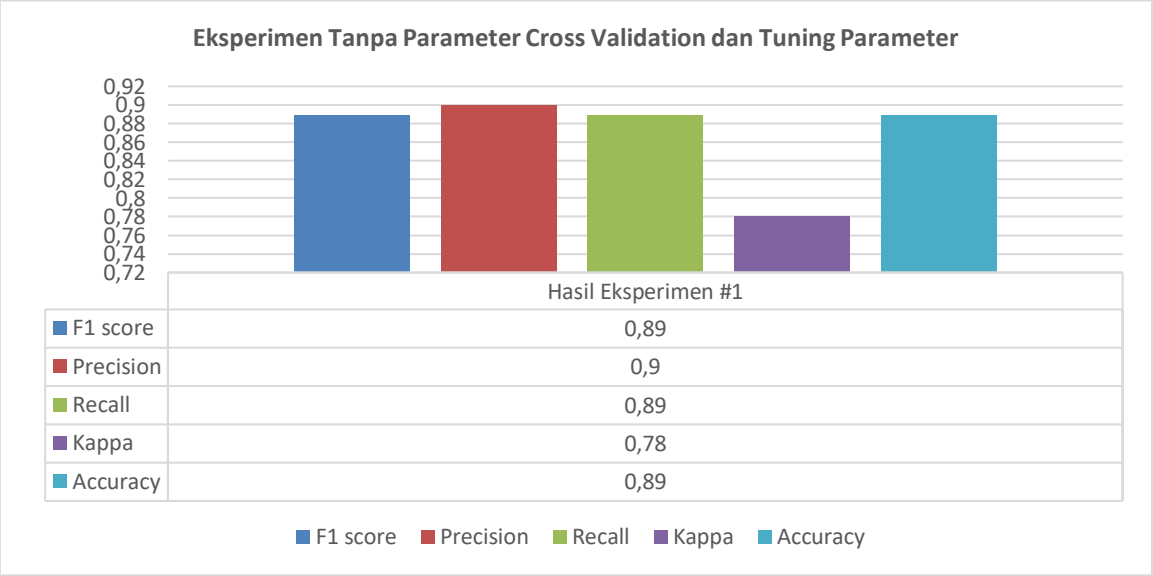
$$y_i(x_i.w + b) \geq 1, \text{untuk } \forall_i$$

Selanjutnya hasil pengujian dilakukan evaluasi untuk menghitung akurasi model, dan dihitung juga perhitungan evaluasi lain seperti *precision*, *recall* dan, *f-measure*.

3. HASIL DAN PEMBAHASAN

Eksperimen riset untuk pra-proses, klasifikasi, validasi dan evaluasi dilakukan dengan bantuan bahasa pemrograman Python dan *scikit-learn library* serta data augmentation dilakukan dengan bantuan R-library. Tahap pra-pemrosesan diselesaikan dengan menggunakan *TfidfVectorizer*. Klasifikasi dalam penelitian ini menggunakan *support vector machine* (SVM) di *librarysklearn*. Validasi dilakukan menggunakan *cross validation* di mana persentase masing-masing sampel pelatihan 70% dan sampel pengujian 30%.

Penelitian dilakukan dalam dua tahap, yaitu eksperimen pertama tanpa parameter *cross validation* dan *tuning parameter*, kemudian dilakukan eksperimen kedua dengan parameter *cross validation* dan *tuning parameter*. Hasil kinerja untuk klasifikasi SVM untuk eksperimen tanpa pparameter cross validation dan tuning parameter dengan menunjukkan detail akurasi, nilai f1, presisi, recall, dan nilai *kappa* dipresentasikan pada **Gambar 2.**



Gambar 2 Eksperimen Tanpa Parameter Cross Validation dan Tuning Parameter

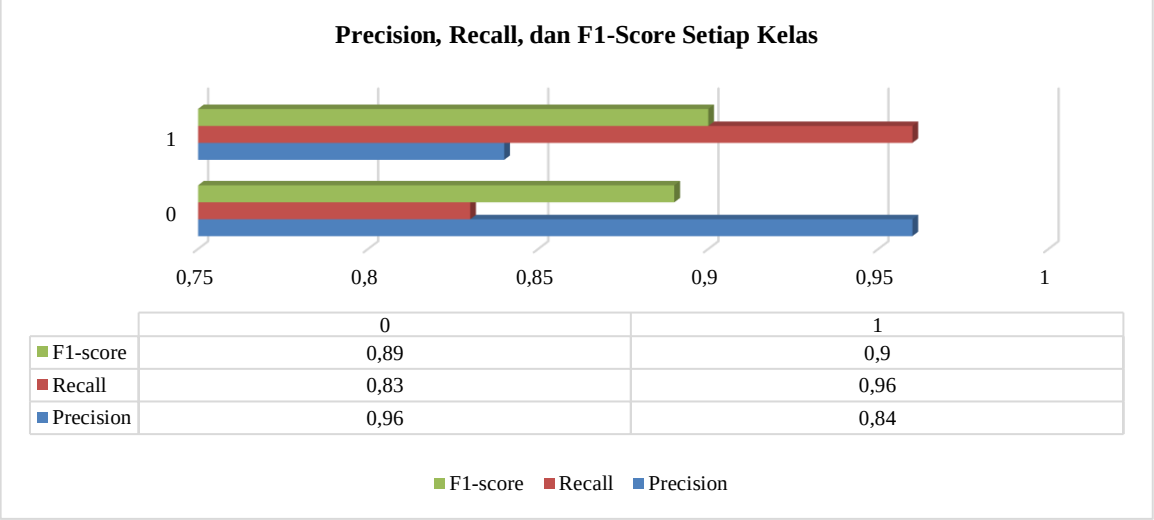
Eksperimen tanpa parameter *cross validation* dan *tuning parameter* menggunakan algoritma SVM mendapatkan akurasi 89%. Nilai akurasi menunjukkan bahwa SVM mampu mengklasifikasikan data teks dengan baik berdasarkan kategori komplain dan bukan komplain. Selain itu, berdasarkan eksperimen didapatkan bahwa nilai *precision* dan *recall* masing-masing adalah 0,90 dan 0,89. Kemudian, untuk menafsirkan nilai dari *cohen's kappa* dapat merujuk ke **Tabel 1**.

Tabel 1 Interpretasi dari Cohen's Kappa

Nilai	*Strength of agreement*
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

Sumber: [36]

Berdasarkan hasil eksperimen, nilai K menggunakan SVM adalah 0,78 dengan waktu komputasi 0.019545 s sehingga dapat dikatakan bahwa *strength of agreement* termasuk dalam kategori 0.61 – 0.80 (baik). Selain itu, nilai *precision*, *recall*, dan *F1-Score* untuk setiap kelas untuk eksperimen tanpa parameter cross validation dan tuning parameter ditunjukkan digambarkan pada **Gambar 3**.



	0	1
F1-score	0,89	0,9
Recall	0,83	0,96
Precision	0,96	0,84

Gambar 3 Eksperimen #1: Precision, Recall dan F1-Score Setiap Kelas

Esperimen kedua dengan parameter cross validation dan tuning parameter. Dalam eksperimen ini digunakan cross validation dengan k-5 dan k-10 dan menerapkan parameter tuning SVM dengan C constant dan nilai gamma. Hasil kinerja untuk klasifikasi SVM untuk eksperimen dengan parameter *cross validation* dan *tuning parameter* ditunjukkan pada **Tabel 2**.

Tabel 2 Hasil Cross Validation

Variabel	k-5	k-10
C constant	32.0	128.0
Gamma	0.000122	3.0517578125e-05
Akurasi	0.9507	0.9648
Waktu komputasi	18.976 s	40.118 s

Berdasarkan eksperimen, *c constant* dan *gamma* diperoleh hasil yang lebih baik untuk *cross validation*. Hasil *cross validation* dengan *k-10* diperoleh akurasi terbaik dengan nilai 96,48% dengan waktu komputasi sebanyak 40,118 detik. Selanjutnya, eksperimen dilakukan untuk menemukan fungsi kernel terbaik di antara *sigmoid*, *linear* dan *radial basis function*. Berdasarkan hasil eksperitmen, fungsi kernel *sigmoid* mencapai akurasi dan waktu komputasi terbaik. Hasil eksperimen dengan detail akurasi dan waktu komputasi untuk setiap fungsi kernel dapat dilihat pada **Tabel 3.**

Tabel 3 Hasil *Fine Tuning*

Kernel	Akurasi	Waktu Komputasi (Detik)
Sigmoid	0.9683	30.90
Linear	0.9666	35.36
Radial Basis Function	0.9648	40.67

4. KESIMPULAN

Adapun kesimpulan dari hasil eksperimen tanpa parameter *cross validation* dan *parameter tuning* dan eksperimen dengan parameter *cross validation* dan *parameter tuning* dilihat berdasarkan nilai akurasi, *precision, recall* dan *F1-Score*. Eksperimen tanpa parameter *cross validation* dan *tuning parameter* untuk support vector machine (SVM) mendapatkan akurasi 89,47% dengan nilai *precision* dan *recall* masing-masing adalah 0,90 dan 0,89. Eksperimen kedua menggunakan *cross validation* dengan *k-5* dan *k-10* dan *parameter tuning* dengan *C constant* dan *gamma value*. Hasil *cross validation* dengan *k-10* diperoleh akurasi terbaik dengan nilai 96,48% denganwaktu komputasi selama 40,118 detik. Selanjutnya, fungsi kernel pada *parameter tuning* yaitu *sigmoid*, *linear* dan *radial basis function* dianalisis dan didapatkan bahwa fungsi kernel *sigmoid* mencapai akurasi dan waktu komputasi terbaik.

UCAPAN TERIMA KASIH

Terima kasih kepada Pusat Penelitian dan Fakultas Ilmu Komputer, Universitas Mercu Buana dan yang telah mendukung pelaksanaan penelitian ini.

REFERENSI

[1] C. Burges, *A tutorial on support vector machines for pattern recognition*. Boston: Kluwer Academic Publishers, 1998.
[2] I. Nurhaida, A. Noviyanto, M. Manurung, and A. M. Arymurthi, "Automatic Indonesian's Batik Pattern Recognition using SIFT Approach," *Procedia Comput. Sci.*, vol. 59, pp. 567–576, 2015.
[3] A. M. Abd and S. M. Abd, "Case Studies in Construction Materials Modelling the strength of lightweight foamed concrete using support vector machine (SVM)," *Case Stud. Constr. Mater.*, vol. 6, pp. 8–15, 2017.
[4] M. Diaby and E. Viennet, "Taxonomy-based Job Recommender Systems On Facebook and LinkedIn Profiles," 2014.
[5] A. Kovacevic, D. Ivanovic, B. Milosavljevic, Z. Konjovic, and D. Surla, "Automatic extraction of metadata from scientific publications for CRIS systems," *Progr. Electron. Libr. Inf. Syst.*, vol. 45, no. 4, pp. 376–396, 2011.
[6] V. N. Vapnik, "An Overview of Statistical Learning Theory," vol. 10, no. 5, pp. 988–999, 1999.
[7] B. Ghaddar and J. Naoum-sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 0, pp. 1–12, 2017.
[8] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," in *Big data analytics-as- a-Service: Architecture, Algorithms, and Applications in Health Informatics in KDD2017*, 2017.
[9] P. Dellia and A. Tjahyanto, "Tax Complaints Classification on Twitter Using Text Mining," *J. Sci.*, vol. 2, no. 1, pp. 11–15, 2017.
[10] P. Roberts and W. Hayes, "Information Needs and the Role of Text Mining in Drug Development," in *Pacific Symposium on Biocomputing*, 2008, pp. 592–603.
[11] J. Wainer and P. Fonseca, "How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms," *Artif. Intell. Rev.*, vol. 54, no. 6, pp. 4771–4797, 2021.
[12] C.-C. Chang and S.-H. Chou, "Tuning of the hyperparameters for L2-loss SVMs with the RBF kernel by the maximum-margin principle and the jackknife technique," *Pattern Recognit.*, vol. 48, no. 12, pp. 3983–3992, 2015.
[13] Y. Ibrahim, E. Okafor, and B. Yahaya, "Optimization of rbf-svm hyperparameters using genetic algorithm for face recognition," *Niger. J. Technol.*, vol. 39, no. 4, pp. 1190–1197, 2021.
[14] H. Noprisson and V. Ayumi, "Implementation of Random Forest for Vehicle Type Classification using Gamma Correction Algorithm," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 3, pp. 444–450, 2023.
[15] H. Noprisson and V. Ayumi, "Railroad Track Damage Detection Using CLAHE-KNN (CLAHE K-

Nearest Neighbor)," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 2, pp. 274–279, 2023.
[16] H. Noprisson, "Enterprise 2.0: Identifying Factors for Technology Adoption Based on Technological, Organizational, Human and Social Dimensions," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 1, pp. 59–64, 2023.
[17] H. Noprisson, "Evaluation of Information System Implementation Support for 6-Area Smart City Development," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 1, pp. 71–76, 2023.
[18] H. Noprisson, "Identification of Success Factor Models for Information Systems Development Projects," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 1, pp. 65–70, 2023.
[19] M. Sadikin and A. Fauzan, "Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 6, pp. 58–64, 2023.
[20] M. Sadikin, D. Ramayanti, and A. P. Indrayanto, "The Graded CNN Technique to Identify Type of Food as The Preliminary Stages to Handle the Issues of Image Content Abundant," in *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, 2020, pp. 108–113.
[21] M. Ramadhani and D. Fitrianah, "Implementation of data mining analysis to determine the tuna fishing zone using DBSCAN algorithm," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 5, pp. 706–711, 2019.
[22] B. Priambodo, N. Ani, and Y. Jumaryad, "Predict Next User Location to Improve Accuracy of Mobile Advertising," *J. Phys. Conf. Ser.*, vol. 1175, no. 1, 2019.
[23] U. Salamah, V. K. Aditya, Y. Jumaryadi, V. Ayumi, and H. Noprisson, "Sistem Penjadwalan Pelayanan Perbaikan Komputer Menggunakan Algoritma Round Robin," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 4, no. 1, pp. 122–131, 2023.
[24] S. Bircher, N. Skou, and Y. Kerr, "Validation of SMOS L1C and L2 Products and Important Parameters of the Retrieval Algorithm in the Skjern River Catchment, Western Denmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2969–2985, 2013.
[25] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Cross Validation Through Two-Dimensional Solution Surface for Cost-Sensitive SVM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1103–1121, 2017.
[26] H. D. Wijaya and W. Gunawan, "Implementation of Analytic Network Process Algorithm in E-Lowker System," *J. Syst. Eng. Inf. Technol.*, vol. 1, no. 1, pp. 18–26, 2022.
[27] W. Gunawan, R. A. Wiradiputra, A. P. Sari, D. Prayama, and E. R. Nainggolan, "Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers Using C 4.5 and Naive Bayes Algorithms," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2145–2153, 2023.
[28] M. Kafai and K. Eshghi, "CROification: Accurate Kernel Classification with the Efficiency of Sparse Linear SVM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 34–48, 2019.
[29] M. Purba *et al.*, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022.
[30] I. H. Ikasari, V. Ayumi, M. I. Fanany, and S. Mulyono, "Multiple regularizations deep learning for paddy growth stages classification from LANDSAT-8," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 512–517.
[31] V. Ayumi, L. M. R. Rere, M. I. Fanany, and A. M. Arymurthy, "Optimization of Convolutional Neural Network using Microcanonical Annealing Algorithm," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 506–511.
[32] V. Ayumi, "Application of Machine Learning for SARS-CoV-2 Outbreak," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 7, no. 5, 2020.
[33] V. Ayumi *et al.*, "Transfer Learning for Medicinal Plant Leaves Recognition: A Comparison with and without a Fine-Tuning Strategy," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022.
[34] I. Nurhaida, V. Ayumi, D. Fitrianah, R. A. M. Zen, H. Noprisson, and H. Wei, "Implementation of deep neural networks (DNN) with batch normalization for batik pattern recognition," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 2045–2053, 2020.
[35] N. Andi, I. Ilham, D. J. Yudha, and F. Jefry, "Analysis of user satisfaction level on cashcloud. Id system with system usability scale method and Spearman's rank correlation," *Int. J. Open Inf. Technol.*, vol. 11, no. 9, pp. 92–99, 2023.
[36] D. . Altman, *Practical Statistics for Medical Students*. London: Chapman and Hall, 1991.