



GA-SVM Wrapper Feature Selection untuk Penanganan Data Berdimensi Tinggi

Ahmad Rifa'i^{1*}, Joko Suntoro², Galet Guntoro Setiaji³

¹Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: rifai@usm.ac.id

²Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: jokosuntoro@usm.ac.id

³Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: gallet@usm.ac.id

ARTICLE INFO

History of the article :

Received 28 February 2024

Received in revised form 13 March 2024

Accepted 25 March 2024

Available online 27 March 2024

Keywords:

Klasifikasi; Seleksi Wrapper; Genetic Algorithm; Support Vector Machine; Data Berdimensi Tinggi

*** Correspondence:**

Telepon:
+62 85640292927

E-mail:
rifai@usm.ac.id

ABSTRACT

The increase in data in recent years has increased significantly due to the use of social media and the transition to the digital era. The techniques for processing data into useful information is called data mining. However, the problem that occurs when applying data mining through classification method is high-dimensional data. Its affects the evaluation results into low value. High-dimensional data is defined as data with a large and complex number of features, the complexity of the features made it difficult to select the optimal subset of features because there are features that are irrelevant. In this study the wrapper method will be used by applying the metaheuristic method, namely the genetic algorithm (GA) for the selection of feature subsets to make it more optimal, and the classifier algorithm used is the Support Vector Machine (SVM) algorithm, this method is called GA-SVM WFS. The result of the GA-SVM WFS method is higher than the SVM method which has averaged accuracy 0,902 and 0,874. In this study, there was a significant difference between the GA-SVM WFS method and the SVM method after the paired t-test has been done, within a p-value of 0,01 with an α value of 0,05.

1. INTRODUCTION

Peningkatan data dalam beberapa tahun terakhir ini mengalami peningkatan yang sangat signifikan karena penggunaan sosial media dan peralihan menjadi era digital. Data tersebut berupa teks, suara, video, musik, dan lain sebagainya [1]. Namun apabila data tersebut tidak diolah dengan benar, maka data tersebut akan menjadi tidak berguna.

Teknik untuk pengolahan data menjadi informasi yang berguna dinamakan dengan data mining [2]. Banyak bidang dan lini bisnis telah menerapkan data mining untuk metode pengolahan data menjadi informasi, seperti: klasifikasi untuk mendeteksi serangan sistem [3], klasifikasi untuk mendeteksi penipuan kartu kredit [4], klasifikasi untuk mendeteksi penyakit kanker payudara [5], klasifikasi untuk prediksi cacat perangkat lunak [6], estimasi biaya dalam supply management [7].

Namun masalah yang terjadi ketika menerapkan data mining, khususnya metode klasifikasi adalah data berdimensi tinggi karena data berdimensi tinggi mempengaruhi hasil evaluasi dalam klasifikasi menjadi rendah [8]. Hal ini disebabkan karena data berdimensi tinggi didefinisikan sebagai data dengan jumlah fitur yang banyak dan kompleks, kompleksitas fitur mengakibatkan sulitnya memilih subset fitur yang optimal karena terdapat fitur yang tidak relevan [9]. Metode untuk pemilihan subset fitur yang optimal dalam data mining disebut dengan seleksi fitur [10].

Teknik yang digunakan untuk seleksi fitur dibagi menjadi tiga kategori yaitu: teknik filter, teknik wrapper, dan teknik hybrid [11]. Teknik filter menggunakan pendekatan intrinsik data untuk pemilihan fitur [12]. Di sisi lain, metode wrapper bekerja berdasarkan evaluasi kinerja *classifier* untuk pemilihan fitur [13]. Sedangkan metode hybrid adalah menggabungkan teknik filter dengan teknik wrapper untuk pemilihan fitur [1].

Teknik yang digunakan untuk seleksi fitur pernah diusulkan oleh beberapa peneliti yaitu teknik filter, teknik wrapper, dan hybrid [14]. Relevansi antar atribut digunakan pada teknik filter berdasarkan sifat intrinsik data [15]. Pada penelitian Solorio, dkk [12] menggunakan teknik filter untuk seleksi fitur pada dataset publik yaitu UCI *Machine Learning Repository*. Teknik filter bertujuan untuk mencari subset fitur yang relevan, metode yang diusulkan Solorio dibagi menjadi dua tahapan yaitu tahapan pembobotan fitur, dan tahapan pemilihan fitur relevan. Setelah diketahui fitur yang relevan dan terbaik, maka langkah selanjutnya dilakukan *classifier* menggunakan algoritma klasifikasi yaitu SVM, k-NN, dan Random Forest. Metode pengukuran evaluasi yang digunakan adalah pengukuran akurasi.

Teknik wrapper menilai subset fitur berdasarkan pada kinerja model klasifikasi. Fitur ditambahkan ataupun dihapus berdasarkan dampak fitur pada akurasi model klasifikasi [16]. Muhammad Shabbir Abbasi menggunakan teknik wrapper untuk klasifikasi *malware* ransomware [17]. Deteksi ransomware digunakan strategi berbasis grup melalui metode *Particle Swarm Optimization* (PSO). Metode PSO bertujuan pemilihan fitur secara otomatis dengan pertimbangan relevansi dan hasil klasifikasi pada *classifier*. Sehingga pada pemilihan fitur tidak dibutuhkan inputan dari manusia, melainkan otomatis menggunakan metode metaheuristik yaitu PSO.

Teknik hybrid merupakan teknik yang menggabungkan teknik filter dan teknik wrapper untuk menghasilkan hasil yang unggul pada kinerja model [14]. Penelitian Amukta Malyada Vommi menggunakan teknik hybrid untuk klasifikasi penyakit covid-19 [18]. Dan penelitian Shivani Jain menggunakan teknik hybrid untuk deteksi kode yang berbau (*smell code*) [19]. Kode berbau didefinisikan kelemahan dalam desain sistem sehingga dimungkinkan akan memperlambat proses development dan dapat menghasilkan bug pada kemudian hari.

Teknik wrapper dengan diterapkan metode metaheuristik menghasilkan hasil evaluasi paling optimal dibandingkan dengan teknik filter dan teknik hybrid [20]. Sehingga dalam penelitian ini akan digunakan teknik wrapper dengan menerapkan metode metaheuristik yaitu algoritma genetika (GA) untuk seleksi fitur, dan algoritma pengklasifikasi yang digunakan adalah algoritma *Support Vector Machine* (SVM).

RESEARCH METHODS

Menurut jenis metodenya, metode penelitian eksperimen akan digunakan dalam penelitian ini [21]. Metode penelitian eksperimen adalah investigasi hubungan kasual akan dilakukan pengujian yang dikontrol oleh peneliti itu sendiri (hubungan sebab-akibat). Langkah-langkah

penelitian ini adalah sebagai berikut: analisis permasalahan dan tinjauan pustaka, pengumpulan dataset, metode yang diusulkan, eksperimen dan pengujian hasil, kemudian evaluasi hasil.

Analisis Permasalahan dan Tinjauan Pustaka

Penelitian ini diawali dengan melakukan pencarian *literature review* untuk melihat *research trends* pada topik data berdimensi tinggi. Kemudian dilakukan pencarian *survey paper* untuk menemukan *state of the art* metode yang cocok untuk digunakan dalam mengatasi permasalahan pada data berdimensi tinggi yang mempunyai banyak fitur. Pada proses analisis permasalahan data berdimensi tinggi didapat kelebihan dan kekurangan dari masing-masing metode yang telah diidentifikasi. Hasil identifikasi tersebut kemudian dijadikan landasan untuk menggunakan metode yang paling baik dari beberapa metode yang ada.

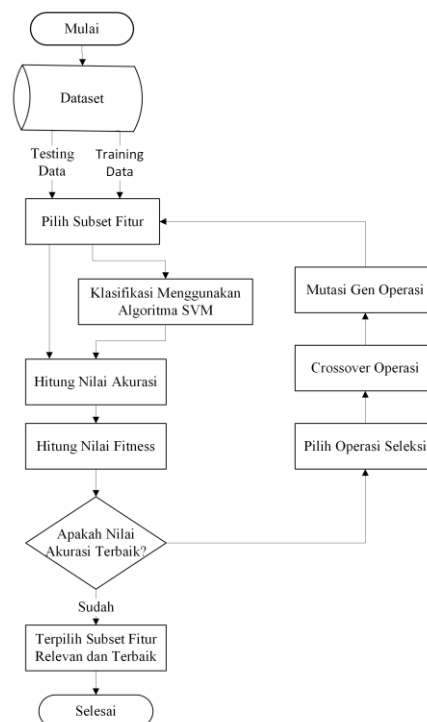
Pengumpulan Dataset

Penggunaan *dataset* pada penelitian ini adalah *dataset* publik yang berasal UCI *Machine Learning Repository*. Tabel 1 menunjukkan dataset pada UCI *Machine Learning Repository* yang mengandung data berdimensi tinggi. Dataset tersebut juga digunakan oleh para peneliti dengan topik data berdimensi tinggi.

Tabel 1. *Dataset* Berdimensi Tinggi

No	Dataset	Jumlah Atribut	Jumlah Data
1	musk	167	476
2	qsar	1.025	1.687
3	sonar	61	208
4	spam	58	4.601
5	vehicle	19	846

Metode Yang Diusulkan



Gambar 1. Metode Yang Diusulkan (GA-SVM WFS)

Dataset yang digunakan mengandung data berdimensi tinggi dengan banyak fitur dan fitur tidak relevan. Metode *wrapper* dengan menerapkan algoritma GA untuk pemilihan fitur optimal yang relevan dan diterapkan algoritma SVM sebagai *classifier*. Metode yang diusulkan ini diberi nama Genetic Algorithm-Support Vector Machine Wrapper Feature Selection (GA-SVM WSF). Gambar 1 menunjukkan flowchart dari metode GA-SVM WFS. Langkah-langkah metode GA-SVM WSF adalah sebagai berikut:

1. Memilih dataset yang mengandung data berdimensi tinggi.
2. Pisahkan data testing dan data training,
3. Kemudian memilih subset fitur.
4. Dataset dengan subset fitur baru dilakukan klasifikasi menggunakan algoritma SVM.
5. Setelah dilakukan klasifikasi lalu hitung nilai akurasi.
6. Hitung nilai fitness, apabila hasil akurasi belum optimal maka dilakukan pemilihan subset fitur dengan cara pilih operasi seleksi, crossover operasi, dan mutasi gen. Apabila hasil akurasi terbaik, maka terpilih subset fitur relevan dan terbaik.

Eksperimen dan Pengujian Metode

Tahapan eksperimen pada penelitian ini adalah sebagai berikut:

1. Menyiapkan dan mengumpulkan dataset KEEL Repository dan UCI Machine Learning Repository yang mengandung data berdimensi tinggi.
2. *Dataset* pada tahapan eksperimen pada penelitian akan menggunakan data training dan data testing menggunakan metode 10-fold cross validation.
3. Melakukan pengujian data training dan data testing dengan menggunakan algoritme Support Vector Machine, kemudian catat hasil akurasinya, metode ini disebut dengan SVM.
4. Melakukan pengujian data training dan data testing dengan menggunakan Genetik Algoritma untuk pemilihan fitur yang optimal kemudian dilakukan classifier menggunakan algoritma Support Vector Machine, kemudian catat hasil evaluasi, metode ini disebut dengan GA-SVM WFS.
5. Membandingkan hasil akurasi pada metode SVM, dan GA-SVM WFS, kemudian mengambil hasil yang terbaik.

Dalam penelitian ini menggunakan program bantu Python versi 3.0, XLSTAT 2016. Sedangkan komputer yang digunakan dalam penelitian ini memiliki spesifikasi yang dapat dilihat pada Tabel 2.

Tabel 2. Spesifikasi Komputer yang Digunakan

Jenis	Keterangan
Processor	Intel Core i7-7500U CPU @ 2.70GHz (4 CPUs)
Memory	16 GB
Hardisk	SSD 1 TB
Sistem Operasi	Windows 10 Enterprise 64-bit
Aplikasi	Python versi 3.0, XLSTAT 2016

Evaluasi Hasil

Confusion matrix adalah tabel yang berisi matriks dua dimensi, salah satu dimensi menunjukkan nilai prediksi dari klasifikasi dan dimensi lainnya menunjukkan nilai aktual dari klasifikasi [22]. Tabel 3 menunjukkan confusion matrix dengan dua kelas [23]. Jika nilai prediksi benar dan nilai aktual benar maka disebut *True Positive* (TP). Jika nilai prediksi salah dan nilai aktual salah maka disebut *True Negative* (TN). Jika nilai prediksi benar dan nilai aktual salah maka

disebut *False Positive* (FP). Jika nilai prediksi salah dan nilai aktual benar maka disebut *False Negative* (FN).

Setelah nilai dari *confusion matrix* diketahui maka langkah selanjutnya adalah menghitung nilai evaluasi yaitu akurasi. Formula yang digunakan untuk menghitung nilai akurasi [24] dapat dilihat pada persamaan 1.

Tabel 3. Confusion Matrix dengan Dua Kelas

Keterangan	Prediksi Benar	Prediksi Salah
Aktual Benar	TP	FN
Aktual Salah	FP	TN

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

Untuk mengetahui apakah terdapat perbedaan yang signifikan antara dua model digunakan salah satu uji statistik yaitu *paired t-test* [25]. *Paired t-Test* membahas perbedaan antara nilai-nilai yang dipasangkan dalam dua sampel, kemudian menghitung variasi nilai dalam setiap sampel dan menghasilkan nilai *t-value*. Persamaan 2 menunjukkan formula penghitungan nilai *t*. Dimana nilai *AVG1* dan *AVG2* adalah nilai rata-rata masing-masing model. *SD* adalah nilai standard deviasi dari dua varian sampel, dan *N* adalah dua ukuran sampel. Pada penelitian ini nilai α yang digunakan adalah 0,05.

$$t = \frac{(AVG1-AVG2)}{\sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}} \quad (2)$$

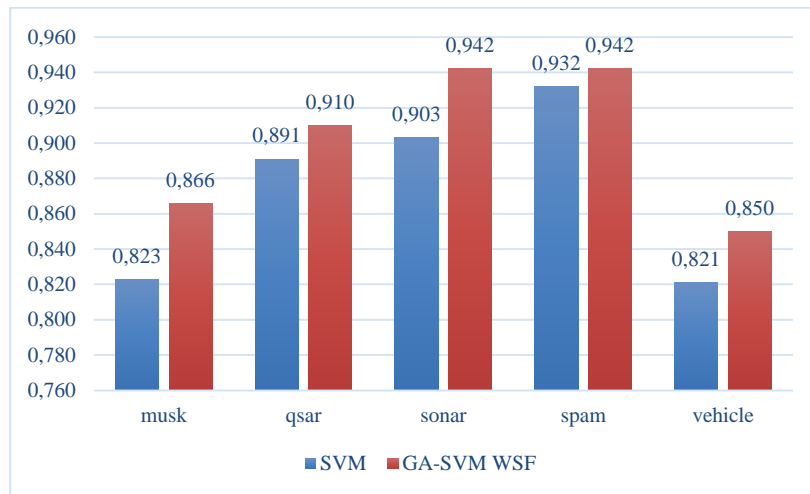
RESULTS

Dalam penelitian ini nilai akurasi dari metode yang diusulkan (GA-SVM WSF) akan dibandingkan dengan metode Support Vector Machine (SVM). Tabel 4 dan Gambar 2 menunjukkan hasil pengukuran akurasi antar metode. Dari pengukuran nilai akurasi antara metode SVM dan GA-SVM WSF didapatkan hasil rata-rata akurasi metode GA-SVM WSF lebih akurat daripada metode SVM dengan nilai rata-rata akurasi metode GA-SVM WSF sebesar 0,902 sedangkan metode SVM sebesar 0,874.

Setelah nilai akurasi pada masing-masing metode didapatkan, maka tahap berikutnya adalah melakukan uji normalitas. Uji normalitas dilakukan dengan tujuan agar diketahui data hasil pengukuran tersebut masuk ke dalam data terdistribusi normal atau masuk ke dalam data tidak terdistribusi normal agar nantinya dapat dipilih uji validasi yang tepat untuk mengetahui apakah terdapat perbedaan secara signifikan antara dua metode yang digunakan.

Tabel 4. Rekap Hasil Pengukuran Akurasi Antar Metode

Datasets	SVM	GA-SVM WSF
musk	0,823	0,866
qsar	0,891	0,910
sonar	0,903	0,942
spam	0,932	0,942
vehicle	0,821	0,850
Rata-Rata	0,874	0,902



Gambar. 1 Diagram Hasil Pengukuran Akurasi Antar Metode

Uji normalitas yang digunakan dalam penelitian ini adalah uji Shapiro-Wilk test. Dalam penelitian ini ditetapkan nilai α sebesar 0,05. Jika nilai $p - value$ lebih besar dari nilai α maka dapat disimpulkan bahwa data terdistribusi normal, begitu sebaliknya jika nilai $p - value$ lebih kecil dari nilai α , maka dapat disimpulkan bahwa data tidak terdistribusi normal. Tabel 5 menunjukkan hasil uji Shapiro-Wilk test pada metode SVM dan Tabel 6 menunjukkan hasil uji Shapiro Wilk test pada metode GA-SVM WFS. Hasil dari uji Shapiro-Wilk test pada metode SVM dan GA-SVM WFS didapatkan nilai $p - value$ lebih besar dari nilai α , maka dapat disimpulkan bahwa kedua metode tersebut data terdistribusi normal.

Dari uji Shapiro-Wilk test didapatkan hasil bahwa data terdistribusi normal, sehingga langkah selanjutnya adalah dilakukan uji validasi yaitu *Paired T-Test* untuk mengetahui apakah terdapat perbedaan signifikan antara metode SVM dan metode GA-SVM WFS. Tabel 7 menunjukkan hasil uji *Paired T-Test* pada penelitian ini, didapatkan nilai $p - value$ lebih kecil dari nilai α yaitu sebesar 0,001. Sehingga dapat disimpulkan bahwa terdapat perbedaan signifikan antara metode SVM dengan metode GA-SVM WFS.

Tabel 5. Hasil Uji Shapiro-Wilk Metode Svm

Keterangan	Nilai
W	0,875
p-value (Two-tailed)	0,287
alpha	0,05

Tabel 6. Hasil Uji Shapiro-Wilk Metode Ga-Svm Wfs

Keterangan	Nilai
W	0,870
p-value (Two-tailed)	0,266
alpha	0,05

Tabel 7. Hasil Uji Paired T-Test

Keterangan	Nilai
Difference	-0,028
t (Observed value)	-4,566
t (Critical value)	2,776
DF	4
p-value (Two-tailed)	0,010
alpha	0,05

DISCUSSION

Hasil penelitian menunjukkan bahwa metode GA-SVM WFS memberikan hasil evaluasi terbaik dibandingkan metode SVM dengan nilai rata-rata akurasi pada metode GA-SVM WFS sebesar 0,902, sedangkan metode SVM sebesar 0,874 ketika digunakan pada dataset berdimensi tinggi yaitu dataset dengan fitur yang banyak. Terdapat perbedaan secara signifikan antara metode GA-SVM WFS dengan metode SVM dengan pengujian *Paired T-Test*.

Penelitian ini telah memberikan kontribusi yaitu pemilihan fitur yang relevan menggunakan teknik wrapper. Untuk pemilihan fitur yang relevan tidak dilakukan secara acak, melainkan diterapkan metode metaheuristik yaitu *Genetic Algorithm Classifier* yang digunakan dalam penelitian ini adalah algoritma *Support Vector Machine*.

CONCLUSIONS AND RECOMMENDATIONS

Pada penelitian selanjutnya mungkin dapat digunakan metode metaheuristik lainnya untuk pemilihan fitur yang relevan secara otomatis, seperti *Ant Colony*, *Bee Colony*, *Firefly Algorithm*, dan lain sebagainya. Selain itu, mungkin dapat digunakan algoritma *classifier* lainnya seperti *Neural Network*, *k-NN*, *Linear Regression*, dan lain sebagainya.

REFERENCES

- [1] J. Wan, H. Chen, Z. Yuan, T. Li, X. Yang, and B. Bin Sang, "A novel hybrid feature selection method considering feature interaction in neighborhood rough set[Formula presented]," *Knowl Based Syst*, vol. 227, Sep. 2021, doi: 10.1016/j.knosys.2021.107167.
- [2] Y. Guo, N. Wang, Z. Y. Xu, and K. Wu, "The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology," *Mech Syst Signal Process*, vol. 142, Aug. 2020, doi: 10.1016/j.ymsp.2020.106630.
- [3] I. F. Kilincer, T. Tuncer, F. Ertam, and A. Sengur, "SPA-IDS: An intelligent intrusion detection system based on vertical mode decomposition and iterative feature selection in computer networks," *Microprocess Microsyst*, vol. 96, p. 104752, Feb. 2023, doi: 10.1016/j.micpro.2022.104752.
- [4] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 1. King Saud bin Abdulaziz University, pp. 145–174, Jan. 01, 2023. doi: 10.1016/j.jksuci.2022.11.008.
- [5] P. S., F. Al-Turjman, and T. Stephan, "An automated breast cancer diagnosis using feature selection and parameter optimization in ANN," *Computers & Electrical Engineering*, vol. 90, p. 106958, Mar. 2021, doi: 10.1016/j.compeleceng.2020.106958.

- [6] K. Thirumorthy and J. J. B. J., “A feature selection model for software defect prediction using binary Rao optimization algorithm,” *Appl Soft Comput*, vol. 131, p. 109737, Dec. 2022, doi: 10.1016/j.asoc.2022.109737.
- [7] F. Bodendorf, P. Merkl, and J. Franke, “Intelligent cost estimation by machine learning in supply management: A structured literature review,” *Comput Ind Eng*, vol. 160, p. 107601, Oct. 2021, doi: 10.1016/j.cie.2021.107601.
- [8] P. Qiu and Z. Niu, “TCIC_FS: Total correlation information coefficient-based feature selection method for high-dimensional data,” *Knowl Based Syst*, vol. 231, p. 107418, Nov. 2021, doi: 10.1016/j.knosys.2021.107418.
- [9] M. García-Torres, R. Ruiz, and F. Divina, “Evolutionary feature selection on high dimensional data using a search space reduction approach,” *Eng Appl Artif Intell*, vol. 117, p. 105556, Jan. 2023, doi: 10.1016/j.engappai.2022.105556.
- [10] B. Wang *et al.*, “Selective Feature Bagging of one-class classifiers for novelty detection in high-dimensional data,” *Eng Appl Artif Intell*, vol. 120, p. 105825, Apr. 2023, doi: 10.1016/j.engappai.2023.105825.
- [11] G. Manikandan and S. Abirami, “An efficient feature selection framework based on information theory for high dimensional data,” *Appl Soft Comput*, vol. 111, p. 107729, Nov. 2021, doi: 10.1016/j.asoc.2021.107729.
- [12] S. Solorio-Fernández, J. Fco. Martínez-Trinidad, and J. A. Carrasco-Ochoa, “A Supervised Filter Feature Selection method for mixed data based on Spectral Feature Selection and Information-theory redundancy analysis,” *Pattern Recognit Lett*, vol. 138, pp. 321–328, Oct. 2020, doi: 10.1016/j.patrec.2020.07.039.
- [13] O. Tarkhaneh, T. T. Nguyen, and S. Mazaheri, “A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm,” *Inf Sci (N Y)*, vol. 565, pp. 278–305, Jul. 2021, doi: 10.1016/j.ins.2021.02.061.
- [14] A. Got, A. Moussaoui, and D. Zouache, “Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach,” *Expert Syst Appl*, vol. 183, p. 115312, Nov. 2021, doi: 10.1016/j.eswa.2021.115312.
- [15] W. BinSaedan and S. Alramlawi, “CS-BPSO: Hybrid feature selection based on chi-square and binary PSO algorithm for Arabic email authorship analysis,” *Knowl Based Syst*, vol. 227, p. 107224, Sep. 2021, doi: 10.1016/j.knosys.2021.107224.
- [16] M. R. Alnowami, F. A. Abolaban, and E. Taha, “A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer,” *J Radiat Res Appl Sci*, vol. 15, no. 1, pp. 104–110, Mar. 2022, doi: 10.1016/j.jrras.2022.01.003.
- [17] M. S. Abbasi, H. Al-Sahaf, M. Mansoori, and I. Welch, “Behavior-based ransomware classification: A particle swarm optimization wrapper-based approach for feature selection,” *Appl Soft Comput*, vol. 121, p. 108744, May 2022, doi: 10.1016/j.asoc.2022.108744.
- [18] A. M. Vommi and T. K. Battula, “A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study,” *Expert Syst Appl*, vol. 218, p. 119612, May 2023, doi: 10.1016/j.eswa.2023.119612.
- [19] S. Jain and A. Saha, “Improving performance with hybrid feature selection and ensemble machine learning techniques for code smell detection,” *Sci Comput Program*, vol. 212, p. 102713, Dec. 2021, doi: 10.1016/j.scico.2021.102713.
- [20] R. Espinosa, F. Jiménez, and J. Palma, “Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data,” *Inf Sci (N Y)*, vol. 622, pp. 1064–1091, Apr. 2023, doi: 10.1016/j.ins.2022.12.004.

-
- [21] J. Suntoro, A. Ilham, and H. A. D. Rani, "New Method Based Pre-Processing to Tackle Missing and High Dimensional Data of CRISP-DM Approach," *J Phys Conf Ser*, vol. 1471, no. 1, p. 012012, Feb. 2020, doi: 10.1088/1742-6596/1471/1/012012.
- [22] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowl Based Syst*, vol. 247, p. 108651, Jul. 2022, doi: 10.1016/j.knosys.2022.108651.
- [23] E. Mortaz, "Imbalance accuracy metric for model selection in multi-class imbalance classification problems," *Knowl Based Syst*, vol. 210, p. 106490, Dec. 2020, doi: 10.1016/j.knosys.2020.106490.
- [24] S. P. Potharaju, M. Sreedevi, V. K. Ande, and R. K. Tirandasu, "Data mining approach for accelerating the classification accuracy of cardiocography," *Clin Epidemiol Glob Health*, vol. 7, no. 2, pp. 160–164, Jun. 2019, doi: 10.1016/j.cegh.2018.03.004.
- [25] N. Kim, "The limit distribution of a modified Shapiro–Wilk statistic for normality to Type II censored data," *J Korean Stat Soc*, vol. 40, no. 3, pp. 257–266, Sep. 2011, doi: 10.1016/J.JKSS.2010.10.004.