Model Optimasi SVM Dengan PSO-GA dan SMOTE Dalam Menangani High Dimensional dan Imbalance Data Banjir

Raenald Syaputra[1], Taghfirul Azhima Yoga Siswa [2*], Wawan Joko Pranoto [3]

[1,2,3]Program Studi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Kalimantan Timur
Email: [1]2011102441040@umkt.ac.id, [2*]tay758@umkt.ac.id, [3]wjp337@umkt.ac.id

(Naskah masuk: 3 Jun 2024, direvisi: 24 Jun 2024, diterima: 25 Jun 2024)

Abstrak

Banjir merupakan salah satu bencana alam yang sering terjadi di Indonesia, termasuk di Kota Samarinda dengan 18-33 titik desa terdampak dari tahun 2018-2021. Penggunaan *machine learning* dalam mengklasifikasi bencana banjir sangat penting untuk memprediksi kejadian di masa mendatang. Beberapa penelitian sebelumnya terkait klasifikasi data banjir dalam 3 tahun terakhir telah dilakukan. Namun, dari beberapa penelitian tersebut memunculkan masalah terkait dengan *dataset high dimensional* yang dapat menurunkan performa model klasifikasi dan menyebabkan *overfitting*. Selain itu, masalah lain juga muncul dalam hal *imbalance data* yang menyebabkan bias terhadap kelas mayoritas dan representasi yang tidak akurat. Oleh karena itu, permasalahan dataset *high dimensional* dan *imbalance data* merupakan tantangan spesifik yang harus diatas dalam klasifkasi data banjir Kota Samarinda. Penelitian ini bertujuan mengidentifkasi fitur-fitur yang diperoleh dari seleksi fitur *Genetic Algorithm* (GA) yang memiliki pengaruh terhadap akurasi klasifikasi data banjir Kota Samarinda menggunakan algoritma *Support Vector Machine* (SVM), serta meningkatkan akurasi klasifikasi data banjir di Kota Samarinda dengan mengimplementasikan algoritma SVM yang dikombinasikan dengan metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk *oversampling*, seleksi fitur dengan GA dan optimasi menggunakan *Particle Swarm Optimization* (PSO). Teknik validasi yang digunakan adalah *10-fold cross validation* dan evaluasi performa menggunakan *confusion matrix*. Data yang digunakan berasal dari BPBD (Badan Penanggulangan Bencana Daerah) dan BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) Kota Samarinda pada tahun 2021-2023 terdiri dari 11 fitur dan 1.095 record. Hasil penelitian menunjukkan bahwa fitur-fitur penting yang terpilih melalui GA adalah temperatur maksimum, kecepatan angin maksimum, arah angin maksimum, arah angin terbanyak, lamanya penyinaran matahari dan kecepatan angin rata-rata. Dengan kombinasi metode SVM, SMOTE, GA dan PSO, akurasi klasifikasi data banjir mencapai 82,28%. Namun, penelitian ini juga menghadapi tantangan seperti kontradiksi hasil dengan penelitian lain terkait penggunaan SMOTE dan variasi hasil akibat karakteristik dataset serta metode pembagian data yang berbeda. Hasil penelitian ini dapat digunakan oleh pemerintah daerah dan badan penanggulangan bencana daerah Kota Samarinda untuk memprediksi kejadian banjir dengan lebih akurat, serta memungkinkan tindakan pencegahan yang lebih efektif. Penerapan hasil penelitian ini dapat meningkatkan efektivitas dalam mitigasi bencana banjir Kota Samarinda.

Kata Kunci: Klasifikasi Banjir, SVM, SMOTE, GA, PSO.

SVM Optimization Model with PSO-GA and SMOTE in Handling High Dimensional and Imbalanced Flood Data

Abstract

Flooding is one of the natural disasters that frequently occurs in Indonesia, including in Samarinda City, with 18-33 affected village points from 2018-2021. The use of machine learning in classifying flood disasters is crucial for predicting future events. Several studies related to flood data classification in the past three years have been conducted. However, these studies have highlighted issues related to high-dimensional datasets that can decrease the performance of classification models and cause overfitting. Additionally, problems related to data imbalance can lead to bias towards the majority class and inaccurate representation. Therefore, the issues of high-dimensional datasets and data imbalance are specific challenges that must be addressed in the classification of flood data in Samarinda City. This study aims to identify important features selected by GA that influence the accuracy of flood data classification in Samarinda City using the SVM algorithm, as well as to improve the

accuracy of flood data classification by implementing the SVM algorithm combined with the Synthetic Minority Oversampling Technique (SMOTE) for oversampling, feature selection with Genetic Algorithm (GA), and optimization using Particle Swarm Optimization (PSO). The validation technique used is 10-fold cross-validation, and performance evaluation is conducted using a confusion matrix. The data used is from BPBD (Regional Disaster Management Agency) and BMKG (Meteorology, Climatology, and Geophysics Agency) Samarinda City from 2021-2023, consisting of 11 features and 1,095 records. The study results show that the important features selected by GA include maximum temperature, maximum wind speed, maximum wind direction, most frequent wind direction, duration of sunshine, and average wind speed. With the combination of SVM, SMOTE, GA, and PSO methods, the flood data classification accuracy reached 82.28%. However, this study also faced challenges such as contradictory results with other studies regarding the use of SMOTE and variations in results due to dataset characteristics and different data partitioning methods. The findings of this study can be used by local governments and disaster management agencies in Samarinda City to predict flood events more accurately, enabling more effective preventive measures. Implementing the results of this study can enhance the effectiveness of flood disaster mitigation in Samarinda City.

Keywords: *Flood Classification, SVM, SMOTE, GA, PSO.*

I. PENDAHULUAN

Indonesia sering mengalami bencana alam akibat iklim tropisnya. Menurut Badan Nasional Penanggulangan Bencana (BNPB), terdapat 8.462 bencana alam pada tahun 2022-2023. Banjir menempati peringkat pertama pada tahun 2022 dan peringkat kedua pada tahun 2023 sebagai bencana yang sering terjadi [1], [2]. Kalimantan Timur, Kota Samarinda sering dilanda banjir, dengan peningkatan kejadian di 18-33 desa dari 2018-2021 [3]. Selain itu, menurut data dari BPBD (Badan Penanggulangan Bencana Daerah) dan BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) Samarinda periode 2021-2023 terdapat 49 kejadian banjir di Samarinda. Oleh karena itu, diperlukannya sistem prediksi banjir menggunakan *machine learning* untuk meningkatkan efektivitas kinerja klasifikasi dalam memprediksi akurasi deteksi frekuensi dan besaran banjir di Samarinda [4].

Berbagai algoritma *machine learning* telah digunakan dalam klasifikasi banjir, seperti *Random Forest* [5], *Naïve Bayes* [6], *K-Nearest Neighbor (KNN)* [7], dan *Support Vector Machine (SVM)* [8]. Rata-rata akurasi yang didapatkan mencapai angka diatas 90% pada dataset *low dimensional*. Dataset *low dimensional* rentan terhadap kehilangan informasi penting, risiko *overfitting*, dan sulit diinterpretasikan [9].

Dimensi dataset mempengaruhi kinerja algoritma. Salah satu masalah dimensi pada dataset yang sering ditemukan adalah *overfitting* dan representasi yang tidak akurat [10]. Seperti dataset *high dimensional* yang memiliki banyak fitur dan sering menyebabkan penurunan performa algoritma, dikarenakan kompleksitas perhitungan, kecenderungan *overfitting*, dan kesulitan dalam visualisasi data [9]. Beberapa penelitian mengenai klasifikasi pada dataset *high dimensional* pernah dilakukan, mulai dari penggunaan algoritma SVM, *Logistic Regression*, *Decision Tree*, KNN, ANN, *Gradient Boosting* dan *Naive Bayes* memberikan performa yang kurang optimal dengan akurasi rendah [11]–[13]. Oleh karena itu, seleksi fitur diperlukan untuk mengurangi jumlah fitur dan mempertahankan yang paling relevan.

Imbalance data sering kali menjadi masalah dalam klasifikasi, menyebabkan *overfitting*, bias terhadap kelas mayoritas, dan representasi yang tidak akurat [14]. Beberapa Penelitian sebelumnya ketika membandingkan performa model seperti *Naive Bayes*, SVM, KNN dan *Decision Tree* antara sebelum dan sesudah teknik *balancing* menunjukkan bahwa *imbalance data* memiliki pengaruh terhadap hasil akhir dari suatu klasifikasi [15], [16]. Oleh karena itu, diperlukannya teknik *oversampling* atau *undersampling* untuk mengatasi *imbalance data*.

Berkaitan dengan klasifikasi data banjir Kota Samarinda, penelitian ini menggunakan data dari BPBD dan BMKG Samarinda periode tahun 2021-2023 dengan total jumlah 21 fitur dan adanya ketimpangan kelas antara kelas terjadi banjir sebesar 49 data dan kelas tidak terjadi banjir sebesar 841 data, hal ini menjadi indikasi bahwa terdapat masalah dataset *high dimensional* dan *imbalance data* pada klasifikasi banjir Kota Samarinda. Masalah dataset *high dimensional* dapat menyebabkan *overfitting* dan kesulitan dalam interpretasi model, sementara *imbalance data* dapat mengakibatkan bias pada model klasifikasi, di mana model cenderung lebih akurat dalam memprediksi kelas mayoritas dibandingkan kelas minoritas [9], [14]. Hal ini dapat mengurangi keakuratan dan keandalan prediksi banjir di Kota Samarinda.

Berdasarkan permasalahan yang terdapat pada klasifikasi data banjir Kota Samarinda, penelitian ini akan menggunakan algoritma *Support Vector Machine* (SVM) untuk klasifikasi data banjir Kota Samarinda. SVM beberapa kali digunakan dalam mengklasifikasi data banjir [12], [13], [17]. SVM terbukti memiliki performa terbaik dibandingkan dengan KNN dan LDA dengan akurasi 97,4% [18]. Namun, SVM memiliki performa yang buruk ketika berhadapan pada dataset *high dimensional* dengan akurasi sebesar 52%-60% [12], [13]. Kelemahan SVM dalam mengklasifikasi data banjir di Kota Samarinda yang memiliki karakteristik dataset *high dimensional* perlu diatasi untuk meningkatkan akurasinya. Untuk itu, diperlukan seleksi fitur guna mengatasi masalah dataset *high dimensional* tersebut.

Seleksi fitur sering kali diterapkan untuk mengatasi dataset *high dimensional* pada penelitian sebelumnya seperti *Genetic Algorithm* (GA), *Gain Ratio*, dan *Information Gain*, terbukti

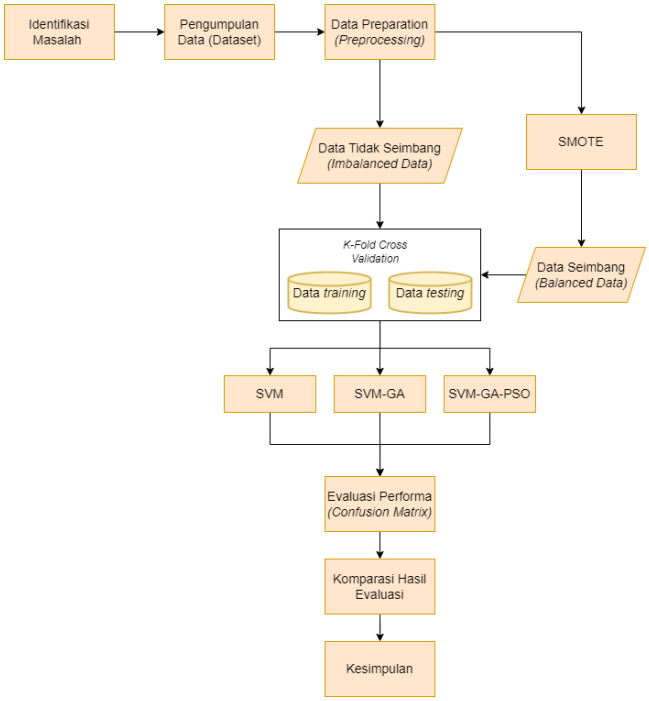
dapat meningkatkan akurasi 3%-13% [13], [19], [20]. Oleh karena itu, penelitian ini akan menerapkan seleksi fitur berupa GA untuk menangani kelemahan SVM sebelumnya. SVM beberapa kali dikombinasikan dengan GA dan memberikan peningkatan akurasi mulai dari 1% hingga 13% [13], [21], [22]. Selain itu, teknik *oversampling* SMOTE akan digunakan untuk menangani *imbalanced data*, yang terbukti efektif dalam meningkatkan akurasi klasifikasi setelah diuji dengan beberapa algoritma seperti *Bayesian Network*, *Decision Tree*, KNN dan *SVM* [15] . Penelitian ini juga akan menerapkan metode optimasi dengan *Particle Swarm Optimization (PSO)* untuk lebih meningkatkan performa SVM, sesuai dengan hasil penelitian sebelumnya yang menunjukkan peningkatan akurasi sebesar 3-11% [23]–[25].

Berdasarkan studi literatur yang telah dilakukan, belum ada peneliti yang mencoba kombinasi algoritma SVM dengan metode seleksi fitur GA, teknik *oversampling* SMOTE dan metode optimasi PSO untuk lebih meningkatkan performa SVM dalam menangani masalah klasifikasi data banjir Kota Samarinda pada dataset *high dimensional* dan *imbalance data*. Oleh karena itu, penelitian ini bersifat baru dan diharapkan dapat memberikan kontribusi dalam meningkatkan akurasi klasifikasi dengan mengatasi kendala yang ada pada dataset *high dimensional* dan *imbalance data* pada data banjir Kota Samarinda.

Penelitian ini bertujuan untuk mengidentifikasi fitur-fitur yang diperoleh dari seleksi fitur GA yang memiliki pengaruh terhadap akurasi klasifikasi data banjir di Kota Samarinda menggunakan algoritma SVM. Selain itu, penelitian ini ingin mengetahui seberapa besar peningkatan akurasi yang dapat dicapai oleh SVM dalam mengklasifikasi data banjir dengan menerapkan metode SMOTE, GA dan PSO. Implementasi algoritma SVM yang dikombinasikan dengan metode-metode tersebut diharapkan dapat mengatasi masalah dataset *high dimensional* serta *imbalance data*, sehingga memberikan performa klasifikasi yang lebih baik. Penelitian ini juga akan mengevaluasi hasil kinerja SVM beserta kombinasinya dengan membagi dataset banjir menggunakan teknik *10-Fold Cross Validation* dan mengukur akurasi menggunakan *confusion matrix*.

II. METODE PENELITIAN

Basis penelitian ini dilakukan dengan metode eksperimen. Metode penelitian eksperimen adalah metode kuantitatif di mana peneliti secara objektif dan sistematis memanipulasi satu atau lebih variabel bebas, mengontrol variabel lain yang relevan, dan mengamati efeknya pada variabel terikat untuk menemukan solusi atas suatu masalah [26], [27]. Eksperimen yang dilakukan dalam penelitian ini yaitu membandingkan kombinasi SVM, GA dan PSO dalam versi SMOTE dan tanpa SMOTE untuk mengklasifikasi data banjir Kota Samarinda. Adapun beberapa tahapan penelitian seperti ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

A. Identifikasi Masalah

Identifikasi masalah akan memandu seluruh proses penelitian. Penelitian ini fokus pada menentukan metode terbaik untuk mengklasifikasi data banjir di Kota Samarinda. Selain itu, dilakukan studi pustaka untuk mengidentifikasi kesenjangan dalam klasifikasi data banjir yang ada.

B. Pengumpulan Data

Penelitian ini menggunakan data sekunder banjir Kota Samarinda dari BPBD (Badan Penanggulangan Bencana Daerah) dan BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) periode 2021-2023. Adapun terkait fitur, data BMKG memiliki 11 fitur dan data BPBD memiliki 10 fitur.

C. *Data Pre-Processing*

Data dari BPBD dan BMKG perlu diolah sebelum permodelan untuk memastikan kualitas data. Tahapan penting meliputi data *integration*, data *selection*, data *transformation*, data *cleaning*, dan data *balancing* [28].

D. Pembagian *Data Training* Dan *Data Testing*

Dataset dibagi menjadi *data training* dan *data testing*. *Data training* melatih model, sementara *data testing* menguji kinerjanya. Teknik *K-Fold Cross-Validation*, dengan k=10, digunakan untuk evaluasi. *Dataset* dibagi menjadi 10 bagian, masing-masing bergantian sebagai *data testing* dan *training*. Rata-rata dari 10 percobaan ini bertujuan untuk memberikan penilaian akurat terhadap kinerja model [29].

E. Permodelan

Penelitian ini menggunakan model klasifikasi SVM dengan SMOTE sebagai *oversampling*, GA sebagai seleksi fitur dan PSO sebagai optimasi. Hasil akhir penelitian akan

membandingkan kinerja model sebelum dan sesudah penggunaan SMOTE.

F. Evaluasi

Evaluasi dilakukan dengan mengukur hasil akhir algoritma pada data *training* menggunakan *confusion matrix* sebagai alat utama evaluasi kinerja [30]. Evaluasi yang dilakukan akan memaparkan performa yang dihasilkan yaitu *accuracy*. Berikut adalah formula *accuracy* yang akan digunakan pada penelitian ini:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

Keterangan: TP *(True Positive)*, TN *(True Negative)*, FP *(False Positive)*, FN *(False Negative)*.

III. HASIL DAN PEMBAHASAN

A. Hasil Pengumpulan Data

Penelitian ini menggunakan data banjir dari BPBD dan BMKG periode 2021-2023. Data mencakup 11 fitur dari BMKG dan 10 fitur dari BPBD. Adapun bentuk data yang diperoleh adalah sebagai berikut.

B. Hasil Data *Pre-Processing*

1. *Data Integration*

Data banjir dari BMKG dan BPBD pada Tabel 1 dan Tabel 2 merupakan dua dataset berbeda yang akan digabungkan untuk menciptakan data sekunder yang lebih lengkap. Proses ini menghasilkan satu dataset dengan 20 fitur, setelah menghilangkan fitur tanggal yang duplikat, untuk memastikan kualitas sebelum klasifikasi. Adapun gabungan beberapa fitur tersebut terdapat pada Tabel 3 sebagai berikut.

2. *Data Selection*

Pada tahap data *selection*, fitur-fitur dipilih berdasarkan relevansinya terhadap penyebab banjir. Fitur tidak relevan dihilangkan dan fitur yang tersisa salah satunya akan ditentukan sebagai kelas atau label. Setelah integrasi data, fitur "jenis bencana" dari BPBD dicocokkan dengan data BMKG berdasarkan tanggal dan diubah menjadi "terjadi banjir" sebagai kelas yang berisi 0 dan 1, artinya tidak terjadi banjir dan terjadi banjir. Dataset ini mencakup 1.095 *record*. Adapun bentuk dataset hasil data *selection* ditampilkan pada Tabel 4 sebagai berikut.

Tabel 1. Data BMKG

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
01-01-2023	25	33,4	29	74		7,6	3	60	2	NE
…	…	…	…	…	…	…	…	…	…	…
31-12-2023	24,6	32,4	28,3	84	0	6,9	4	90	2	E

Tabel 2. Data BPBD

NO	TANGGAL	JAM KEJADIAN	JENIS BENCANA	LOKASI/WILAYAH KELURAHAN / KECAMATAN	LUAS AREA M^2	JUMLAH OBYEK YANG TERKENA BENCANA	KORBAN KL	KS	KH	KM	KK	JUMLAH JIWA	KERUGIAN (RP)		KETERANGAN
1	03 Januari 2021	-	Banjir	Jl. Irigasi RT. 50 Kel. Rawa Makmur Kec. Palaran (Dataran Rendah) Wilayah Handil Bakti RT. 1, RT. 2, RT 3 (Dataran Rendah)	-	Jalan mejadi Susah Untuk Di lalui Dan Menganggu aktivitas warga	-	-	-	-	-	-	Rp.	-	Genangan Air Penyebab Air Sungai Mahakam pasang Dan Lokasi Banjir adalah Dataran Rendah
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
14	Selasa, 31 Januari 2023	Pukul 19.45 wita	Pohon Tumbang	Jl. P. Antasari Pondok Wira 1 Kel. Teluk Lerong Ulu	-	Dampak: Menutup Bahu Jalan dan Menggangu Aktifitas Warga Sekitar	-	-	-	-	-	-	-	-	Penyebab terjadinya Hujan dengan intensitas sedang dan angin kencang

Tabel 3. Fitur Hasil Data *Integration*

No	Fitur	Tipe Data
1	Tanggal	*date*
2	Jam Kejadian	*string*
3	Jenis Bencana	*string*
4	Lokasi Wilayah	*string*
5	Luas Area M^2	*string*
6	Objek Terkena Bencana	*string*
7	Korban	*numeric*
8	Jumlah Jiwa	*numeric*
9	Kerugian	*string*

No	Fitur	Tipe Data
10	Keterangan	*string*
11	Temperatur-maksimum (Tn) (°C)	*numeric*
12	Temperatur-minimum (Tx) (°C)	*numeric*
13	Temperatur-rata-rata (Tavg) (°C)	*numeric*
14	Kelembaban-rata-rata (RH_avg) (%)	*numeric*
15	Curah-hujan (RR) (mm)	*numeric*
16	Lamanya-penyinaran-matahari (ss) (hrs)	*numeric*
17	Kecepatan-angin-maksimum (ff_x) (m/s)	*numeric*
18	Arah-angin-maksimum (ddd_x) (°)	*numeric*
19	Kecepatan-angin-rata-rata (ff_avg) (m/s)	*numeric*
20	Arah-angin-terbanyak (ddd_car) (°)	*numeric*

Tabel 4. Hasil *Data Selection*

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car	terjadi banjir
1/1/2021	23	33,2	26,5	88	1,8	3,3	4	280	2	W	0
…	…	…	…	…	…	…	…	…	…	…	…
31/12/2023	24,6	32,4	28,3	84	0	6,9	4	90	2	E	0

3. *Data Transformation*

Pada tahap data *transformation*, format *string* diubah menjadi numerik untuk memudahkan algoritma. Fitur "arah angin terbanyak (ddd_car)" yang berisi simbol arah mata angin diubah menjadi angka menggunakan *LabelEncoder* dari *library sklearn.preprocessing* [31]. Tabel 5 menunjukkan perbedaan sebelum dan sesudah transformasi, dari simbol (*string*) ke bentuk numerik.

Tabel 5. Sebelum dan Sesudah Data *Transformation*

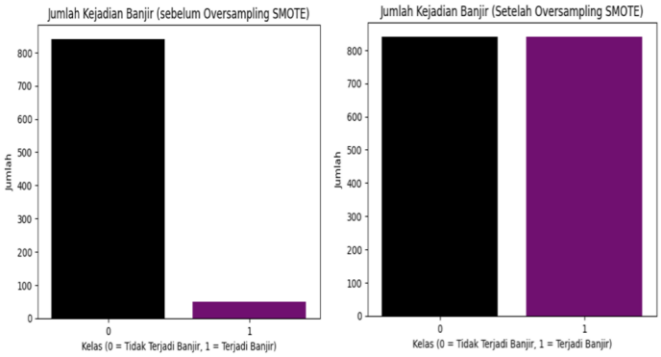
No	(Sebelum) Arah-angin-terbanyak (ddd_car)	(Sesudah) Arah-angin-terbanyak (ddd_car)
1	W	8
…	…	…
1094	E	1

4. *Data Cleaning*

Pada proses data *cleaning*, baris dengan nilai kosong atau nilai *NaN (Not a Number)* dihapus. Dalam proses ini jumlah data dikurangi dari 1.095 *record* menjadi 890 *record*.

5. *Data Balancing*

Tahap data *balancing* merupakan proses menyeimbangkan distribusi kelas dalam dataset untuk menghindari bias pada algoritma klasifikasi akibat ketidakseimbangan jumlah sampel antar kelas. *Imbalance data* dengan 841 untuk kelas 0 (tidak terjadi banjir) dan 49 untuk kelas 1 (terjadi banjir). Teknik SMOTE digunakan untuk menyeimbangkan kelas dengan membuat sampel sintetis untuk kelas minoritas. Dengan Menggunakan modul *python imblearn.over_sampling* dan mengimpor fungsi SMOTE untuk melakukan *oversampling*. Setelah *oversampling*, kedua kelas menjadi seimbang dengan masing-masing 841 *record*, sehingga total data meningkat menjadi 1.682 *record*.



Gambar 2. Sebelum dan Sesudah Teknik *Oversampling*

C. Hasil Pembagian *Data Training* dan *Data Testing*

Pembagian dataset menjadi *data training* dan *testing* sangat penting untuk kinerja model *machine learning*. Menggunakan *10-Fold Cross-Validation*, *dataset* dibagi menjadi sepuluh bagian sama besar. Setiap iterasi, satu bagian digunakan sebagai *data testing* dan sembilan sebagai *data training*, mengurangi bias dan variasi dalam estimasi kinerja model serta memastikan setiap sampel diuji.

D. Hasil Permodelan Dan Evaluasi

Tahap ini akan menampilkan hasil pembelajaran algoritma dalam bentuk akurasi yang dicapai oleh model-model mulai dari SVM beserta model kombinasi lainnya seperti SMOTE, GA dan PSO yang dijelaskan sebelumnya terhadap klasifikasi data banjir, beserta beberapa hasil dari penerapan algoritma seleksi fitur GA terhadap *dataset* yang memberikan kombinasi fitur-fitur terbaiknya.

1. Implementasi SVM

Algoritma SVM diimplementasikan menggunakan *Python* tanpa SMOTE dan dengan SMOTE. Evaluasi akurasi akan dilakukan dengan *confusion matrix* melalui *10-Fold Cross-Validation*. Hasil rata-rata akurasi dan *Confusion Matrix* dirangkum dalam Tabel 6 dan Tabel 7 yang dihitung berdasarkan formula 1.

Tabel 6. Hasil Rata-Rata Akurasi SVM Tanpa SMOTE

TP	FP	TN	FN	*Accuracy*
0	0	841	49	94,49%

$$Accuracy = \frac{0+841}{0+841+0+49} \times 100\% = 94{,}49\%$$

Tabel 7. Hasil Rata-Rata Akurasi SVM Dengan SMOTE

TP	FP	TN	FN	*Accuracy*
653	544	297	188	56,48%

$$Accuracy = \frac{653+297}{653+297+544+188} \times 100\% = 56{,}48\%$$

2. Implementasi SVM-GA

Seleksi fitur GA diimplementasikan menggunakan *Python* tanpa SMOTE dan dengan SMOTE. GA digunakan untuk memilih fitur terbaik berdasarkan nilai *fitness* atau kecocokan kombinasi antar fitur. GA memanggil *library* dari *deap* yang memanggil fungsi *base*, *creator*, *tools* dan *algorithms.* Kemudian mendefinisikan evaluasi SVM sebagai *fitness*, menetapkan tipe individu dan populasi, serta mendefinisikan operasi seleksi, mutasi, dan evaluasi untuk GA. Setelah menetapkan parameter populasi dan generasi, seleksi fitur GA dijalankan untuk menemukan fitur-fitur terbaik. Berikut adalah fitur-fitur terbaik hasil dari seleksi fitur GA.

Tabel 8. Hasil Seleksi Fitur GA Tanpa Dan Dengan SMOTE

Reduksi Fitur GA Tanpa SMOTE	Reduksi Fitur GA Dengan SMOTE
Temperatur maksimum (Tx)	Lamanya penyinaran matahari (ss)
Kecepatan angin maksimum (ff_x)	Kecepatan angin maksimum (ff_x)
Arah angin maksimum (ddd_x)	Kecepatan angin rata-rata (ff_avg)
Arah angin terbanyak (ddd_car)	Arah angin terbanyak (ddd_car)

Hasil seleksi fitur yang didapatkan GA pada Tabel 8 akan diimplementasikan ke algoritma SVM untuk di analisis perubahan yang terjadi. Berikut Tabel 9 dan Tabel 10 yang menampilkan hasil implementasi SVM-GA.

Tabel 9. Hasil Rata-Rata Akurasi SVM-GA Tanpa SMOTE

TP	FP	TN	FN	*Accuracy*
0	0	841	49	94,49%

$$Accuracy = \frac{0+841}{0+841+0+49} \times 100\% = 94{,}49\%$$

Tabel 10. Hasil Rata-Rata Akurasi SVM-GA Dengan SMOTE

TP	FP	TN	FN	*Accuracy*
710	345	496	131	71,70%

$$Accuracy = \frac{710+496}{710+496+345+131} \times 100\% = 71{,}70\%$$

3. Implementasi SVM-GA-PSO

Metode optimasi PSO diimplementasikan menggunakan *Python* tanpa SMOTE dan dengan SMOTE. Metode PSO mengoptimalkan fitur terbaik hasil seleksi fitur GA bersama dengan parameter *C* dan *gamma* pada SVM. PSO diimplementasikan menggunakan *library pyswarm*, menjalankan optimasi dengan menghitung nilai *fitness* dari SVM serta menetapkan batasan nilai *C* dan *gamma*. Dalam proses ini PSO tanpa SMOTE optimasi mendapatkan nilai *C* dan *gamma* sebesar 0,001 dan 0,61, Sedangkan dengan SMOTE mendapatkan nilai *C* dan *gamma* sebesar 99 dan 0,98. Nilai *C* dan *gamma* ini kemudian diterapkan pada proses SVM beserta dengan fitur hasil seleksi fitur dari Tabel 8. Berikut Tabel 11 dan Tabel 12 yang menampilkan hasil implementasi SVM-GA-PSO.

Tabel 11. Hasil Rata-Rata Akurasi SVM-GA-PSO Tanpa SMOTE

TP	FP	TN	FN	*Accuracy*
0	0	841	49	94,49%

$$Accuracy = \frac{0+841}{0+841+0+49} \times 100\% = 94{,}49\%$$

Tabel 12. Hasil Rata-Rata Akurasi SVM-GA-PSO Dengan SMOTE

TP	FP	TN	FN	*Accuracy*
711	168	673	130	82,28%

$$Accuracy = \frac{711+673}{711+673+168+130} \times 100\% = 82{,}28\%$$

E. Pembahasan

Penelitian ini menggunakan GA untuk seleksi fitur pada dataset dengan dan tanpa teknik *oversampling* SMOTE. Tanpa SMOTE, nilai *fitness* mencapai 94,49%, dengan SMOTE nilai *fitness* mencapai 70,93% dan hasil seleksi fiturnya seperti dipaparkan pada tabel 8. Hasil ini menjadi jawaban mengenai kumpulan fitur yang diperoleh seleksi fitur GA. Selain itu, perbedaan hasil seleksi fitur antara dengan dan tanpa teknik *oversampling* SMOTE menunjukkan bagaimana metode seleksi fitur *GA* dapat beradaptasi terhadap perubahan distribusi data akibat teknik *oversampling*.

Penelitian lain mendukung hasil ini. Misalnya, Dilla Evitasari *et al.* [13] menggunakan GA dan teknik *oversampling*, meningkatkan akurasi SVM dari 52,71% menjadi 66,16% dengan tiga fitur terbaik yaitu kelembapan, lamanya penyinaran matahari, dan kecepatan angin maksimum. Kaur dan Bala [32] menggunakan GA tanpa *oversampling*, memilih 10 fitur terbaik dari 15 fitur yaitu tutupan awan, tekanan permukaan laut darwin, suhu kisaran diurnal, temperatur maksimum, temperatur minimum,

evapotranspirasi potensial, curah hujan, kelembapan, frekuensi hari basah dan evapotranspirasi tanaman acuan, mencapai akurasi 88,57% pada SVM. Intan dan Sari [19] menggunakan *Gain Ratio* dan SMOTE, meningkatkan akurasi KNN dari 84% menjadi 89% dengan tiga fitur terbaik yaitu kelembapan, temperatur maksimum, dan temperatur minimum. Adapun detail persamaan hasil seleksi fitur dipaparkan pada Tabel 13.

Persamaan hasil seleksi fitur yang ditunjukkan oleh Tabel 13 memperkuat bahwa fitur-fitur tersebut merupakan elemen penting yang memiliki pengaruh terhadap dataset banjir. Namun, perbedaan hasil seleksi fitur yang bervariasi dapat disebabkan berbagai faktor, termasuk perbedaan dataset, metode seleksi fitur, dan pengolahan data. Metode seperti GA menggunakan fungsi *fitness*, sedangkan *Gain Ratio* menggunakan nilai korelasi, menghasilkan hasil yang berbeda (Tabel 14 dan Tabel 15). Penggunaan *oversampling* juga mempengaruhi hasil. Semua fitur yang diperoleh dalam penelitian ini berpengaruh terhadap data banjir, sebagaimana didukung oleh Tabel 15 yang menunjukkan bahwa seleksi fitur GA cukup berpengaruh terhadap akurasi SVM dalam klasifikasi data banjir Kota Samarinda.

Penelitian ini juga mengevaluasi peningkatan akurasi model dengan dan tanpa *oversampling* SMOTE. Tanpa SMOTE, rata-rata akurasi untuk SVM, SVM-GA, dan SVM-GA-PSO adalah 94,49%, tanpa perubahan akurasi, seperti yang dipaparkan pada tabel 14. Sementara itu model dengan SMOTE, seperti yang dipaparkan pada tabel 15, akurasi meningkat dari 56,48% menjadi 71,70% untuk SVM-GA dan 82,28% untuk SVM-GA-PSO. Hal ini menunjukkan peningkatan 15,22% dari SVM ke SVM-GA, 25,80% dari SVM ke SVM-GA-PSO, dan 10,58% dari SVM-GA ke SVM-GA-PSO. Meskipun akurasi model dengan SMOTE lebih rendah dibandingkan akurasi model tanpa SMOTE, perbedaan ini dapat disebabkan karena model dengan SMOTE dapat mempelajari data yang lebih lengkap dan variatif.

Namun, perbandingan hasil menunjukkan kontradiksi dengan penelitian lain terkait perbedaan antara model SVM dengan dan tanpa penggunaan *oversampling* SMOTE dalam klasifikasi data banjir Kota Samarinda, dari 94,49% menjadi 56,48%. Penelitian lain yang dilakukan oleh Aditya Gumilar *et al.* [5] meningkatkan akurasi SVM dari 78,69% menjadi 90,89% dengan SMOTE, dan Razali *et al.* [15] yang meningkatkan akurasi SVM dari 99,59% menjadi 99,76% dengan SMOTE. Dari sini dapat diketahui bahwa adanya kontradiksi hasil yang terjadi dengan penelitian lain. Perbedaan ini dapat disebabkan oleh beberapa faktor, seperti karakteristik dataset, distribusi data dan metode pembagian data yang berbeda. Seperti penelitian Razali *et al.* [15] yang menemukan bahwa metode SMOTE memang sangat berguna dalam mengatasi data yang tidak seimbang, namun hasilnya tidak selalu memberikan peningkatan akurasi, dalam eksperimen yang mereka lakukan menggunakan model *Bayesian Network* pada dataset banjir Kuala Krai dan memberikan penurunan akurasi setelah penggunaan SMOTE dari 99,94% menjadi 99,68%, lalu dalam kesimpulannya mereka mengungkapkan bahwa efektivitas dari SMOTE dapat bervariasi tergantung pada karakteristik dataset dan pengaturan eksperimen metode yang digunakan. Penelitian Eom *et al.* [33] juga mengungkapkan bahwa SMOTE sering kali tidak efektif dalam menghadapi dataset berdimensi tinggi yang dapat menyebabkan masalah *overfitting*, namun dapat lebih berguna untuk dataset berdimensi rendah.

Tabel 13. Persamaan Hasil Seleksi Fitur Penelitian Ini Dengan Penelitian Lain

Hasil Seleksi Fitur GA Dengan SMOTE Dan Tanpa SMOTE	Penelitian Dilla (SVM-GA)	Penelitian Kaur (SVM-GA)	Penelitian Intan (KNN-Gain Ratio)
Temperatur maksimum (Tx)		✓	✓
Kecepatan angin maksimum (ff_x)			
Arah angin maksimum (ddd_x)	✓		
Arah angin terbanyak (ddd_car)			
Lamanya penyinaran matahari (ss)	✓		
Kecepatan angin rata-rata (ff_avg)			

Tabel 14. Perbandingan Hasil Rata-Rata Akurasi Model Tanpa SMOTE

SVM	SVM-GA	SVM-GA-PSO	Perubahan SVM ke SVM-GA	Perubahan SVM ke SVM-GA-PSO	Perubahan SVM-GA ke SVM-GA-PSO
94,49%	94,49%	94,49%	0%	0%	0%

Tabel 15. Perbandingan Hasil Rata-Rata Akurasi Model Dengan SMOTE

SVM	SVM-GA	SVM-GA-PSO	Perubahan SVM ke SVM-GA	Perubahan SVM ke SVM-GA-PSO	Perubahan SVM-GA ke SVM-GA-PSO
56,48%	71,70%	82,28%	15,22%	25,80%	10,58%

Di samping itu, dari segi metode pembagian data seperti model KNN yang digunakan dalam penelitian Razali *et al.* [15] dengan metode pembagian data *10-fold cross validation* memberikan peningkatan akurasi antara sebelum dan sesudah SMOTE dari 99,50% menjadi 99,76%, lain halnya dengan penelitian yang dilakukan oleh Wibowo *et al.* [34] menggunakan model KNN dengan metode pembagian data *splitting* memberikan penurunan akurasi antara sebelum dan sesudah SMOTE dari 99,96% menjadi 99,76%. Hal ini membuktikan bahwa metode pembagian data yang berbeda juga menjadi faktor dari kontradiksi yang terjadi.

Lebih lanjut, penelitian lain juga ada yang memperkuat hasil penelitian ini berupa peningkatan akurasi dengan algoritma dan kombinasinya seperti Dilla Evitasari *et al.* [13] meningkatkan akurasi SVM dari 52,71% menjadi 66,16% dengan GA, Kanwal *et al.* [35] meningkatkan akurasi SVM dari 81% menjadi 90% dengan GA, Saputra *et al.* [25] meningkatkan akurasi SVM dari 81,59% menjadi 84,81% dengan PSO, lalu pengaruh PSO diperkuat lagi oleh penelitian Faldi *et al.* [36] meningkatkan akurasi Naïve Bayes dari 91,12% menjadi 94,38% dengan PSO, dan Maulidina *et al.* [37] meningkatkan akurasi SVM-GA-PSO dari 90,90% menjadi 97,69%.

Hasil penelitian ini konsisten dengan studi sebelumnya, menunjukkan bahwa kombinasi algoritma seperti SVM-SMOTE, SVM-GA, SVM-PSO, dan SVM-GA-PSO dapat meningkatkan akurasi model. Sehingga, kombinasi SVM, SMOTE, GA, dan PSO merupakan kontribusi unik dari penelitian ini dibandingkan dengan penelitian lain yang hanya menggunakan metode tunggal atau kombinasi yang tidak lengkap. Penelitian ini membuktikan bahwa kombinasi tersebut efektif dalam meningkatkan akurasi model klasifikasi data banjir di Kota Samarinda, mengatasi kelemahan metode tunggal, serta memberikan solusi yang lebih stabil dan adaptif untuk dataset *high dimensional* dan *imbalance data*.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penerapan metode seleksi fitur menggunakan GA pada data banjir Kota Samarinda, fitur penting yang terpilih meliputi temperatur maksimum (Tx), kecepatan angin maksimum (ff_x), arah angin maksimum (ddd_x), arah angin terbanyak (ddd_car), lamanya penyinaran matahari (ss) dan kecepatan angin rata-rata (ff_avg). Peningkatan akurasi klasifikasi data banjir di Kota Samarinda dengan menggunakan kombinasi SVM, SMOTE, GA, dan PSO mencapai 82,28%. Namun, penelitian ini menghadapi kendala seperti kontradiksi hasil dengan penelitian lain terkait penggunaan SMOTE, yang dalam penelitian ini menurunkan akurasi dari 94,49% menjadi 56,48%. Faktor-faktor seperti karakteristik dataset, distribusi data dan metode pembagian data yang berbeda menjadi penyebab dalam kontradiksi ini.

Hasil penelitian ini memiliki manfaat praktis, di mana model yang dikembangkan dapat digunakan oleh pemerintah daerah dan badan penanggulangan bencana daerah Kota Samarinda untuk memprediksi kejadian banjir dengan lebih akurat dan memungkinkan tindakan pencegahan yang lebih efektif. Penelitian ini membuka peluang untuk eksplorasi lebih lanjut terhadap teknik *oversampling* atau *undersampling* lainnya serta metode seleksi fitur dan optimasi lain untuk meningkatkan performa klasifikasi dalam domain yang menghadapi masalah dataset *high dimensional* dan *imbalance data*. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi teoretis tetapi juga aplikasi praktis yang penting dalam bidang mitigasi bencana banjir Kota Samarinda.

B. Saran

Berdasarkan kesimpulan yang didapat pada penelitian ini, terdapat beberapa saran yang dapat menjadi rujukan untuk penelitian selanjutnya. Penelitian selanjutnya dapat menerapkan teknik oversampling lain seperti ADASYN atau teknik undersampling untuk membandingkan efektivitasnya dalam meningkatkan akurasi klasifikasi. Selain itu, eksplorasi lebih lanjut terhadap kernel SVM juga dapat dilakukan untuk melihat peningkatan kinerja klasifikasi. Penelitian mendatang juga dapat mencoba metode seleksi fitur atau optimasi lainnya, seperti metode optimasi Bat Algorithm (BA) untuk mengatasi kebisingan dalam data. Hal ini dapat memberikan wawasan tambahan mengenai apakah ada metode seleksi fitur atau optimasi lain yang lebih efisien dan efektif.

REFERENSI

[1] R. Mustajab, "BNPB: Indonesia Alami 3.522 Bencana Alam pada 2022," *Dataindonesia.id*, 2023.

[2] F. S. Pratiwi, "Data Kejadian Bencana Alam di Indonesia Sepanjang Tahun 2023," *Dataindonesia.id*, 2024.

[3] BPS Kota Samarinda, "Jumlah Desa /Kelurahan yang Mengalami Bencana Alam (Banjir) Menurut Kecamatan di Kota Samarinda." p. https://samarindakota.bps.go.id, 2020.

[4] L. Tarasova *et al.*, "Causative classification of river flood events," *Wiley Interdiscip. Rev. Water*, vol. 6, no. 4, pp. 1–23, 2019, doi: 10.1002/wat2.1353.

[5] Aditya Gumilar, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "Performance Analysis of Hybrid Machine Learning Methods on Imbalanced Data (Rainfall Classification)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3, pp. 481–490, 2022, doi: 10.29207/resti.v6i3.4142.

[6] N. M. Nawi, M. Makhtar, M. Z. Salikon, and Z. A. Afip, "A comparative analysis of classification techniques on predicting flood risk," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 3, pp. 1342–1350, 2020, doi: 10.11591/ijeecs.v18.i3.pp1342-1350.

[7] A. A. Bagaskara, "Klasifikasi Daerah Rawan Banjir menggunakan 10-Fold Cross Validation dan K-Nearest Neighbors," vol. 13, pp. 315–323, 2023, [Online]. Available: https://repository.uksw.edu/handle/123456789/32366%0Ahttps://repository.uksw.edu/bitstream/123456789/32

366/3/T1_682019079_Daftar Pustaka.pdf
[8] T. A. Khan, M. Alam, S. F. Ahmed, Z. Shahid, and M. S. Mazliham, "A Factual Flash Flood Evaluation using SVM and K-NN," *ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci.*, 2019, doi: 10.1109/ICETAS48360.2019.9117424.
[9] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar Joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," *Procedia Comput. Sci.*, vol. 165, pp. 104–111, 2019, doi: 10.1016/j.procs.2020.01.079.
[10] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Inf.*, vol. 12, no. 8, 2021, doi: 10.3390/info12080286.
[11] Sopiatul Ulum, R. F. Alifa, P. Rizkika, and C. Rozikin, "Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum," *Gener. J.*, vol. 7, no. 2, pp. 141–146, 2023, doi: 10.29407/gj.v7i2.20270.
[12] M. R. Ahmmed, J. Monir, and S. A. Khushbu, "Analysis of Flood Risk Prediction Using Different Machine Learning Classifiers: A Study of Predicting Flood Risk in Rural Areas, Bangladesh," *2022 13th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2022*, pp. 1–6, 2022, doi: 10.1109/ICCCNT54827.2022.9984449.
[13] Y. Dilla Evitasari, W. J. Pranoto, and N. Adzmi Verdikha, "Evaluasi Support Vector Machine Dengan Optimasi Metode Genetic Algorithm Pada Klasifikasi Banjir Kota Samarinda Evaluation Support Vector Machine With Optimization Genetic Algorithm Method On Flood Classification In Samarinda," *J. Sains Komput. dan Teknol. Inf.*, vol. 6, no. 1, pp. 49–53, 2023.
[14] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0192-5.
[15] N. Razali, S. Ismail, and A. Mustapha, "Machine learning approach for flood risks prediction," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 73–80, 2020, doi: 10.11591/ijai.v9.i1.pp73-80.
[16] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *Sistemasi*, vol. 10, no. 2, p. 445, 2021, doi: 10.32520/stmsi.v10i2.1303.
[17] D. Fitrianah, W. Gunawan, and A. Puspita Sari, "Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction," *Februari*, vol. 21, no. 1, pp. 1–11, 2022.
[18] T. A. Khan, Z. Shahid, M. Alam, M. M. Su'ud, and K. Kadir, "Early Flood Risk Assessment using Machine Learning: A Comparative study of SVM, Q-SVM, K-NN and LDA," *MACS 2019 - 13th Int. Conf. Math. Actuar. Sci. Comput. Sci. Stat. Proc.*, 2019, doi: 10.1109/MACS48846.2019.9024796.
[19] S. Intan and P. Sari, "Analisis Pengaruh Gain Ratio Untuk Algoritma K-Nearest Neighbor Pada Klasifikasi Data Banjir Di Kota Samarinda Analysis Of The Effect Of Gain Ratio For Algorithms K-Nearest Neighbor On Classsification Flood Data In Samarinda City," vol. 6, no. 1, pp. 54–59, 2023.
[20] M. Norhalimi and T. A. Y. Siswa, "Optimasi Seleksi Fitur Information Gain pada Algoritma Naïve Bayes dan K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 3, pp. 237–255, 2022, doi: 10.14421/jiska.2022.7.3.237-255.
[21] H. Harafani and A. Maulana, "Penerapan Algoritma Genetika pada Support Vector Machine Sebagai Pengoptimasi Parameter untuk Memprediksi Kesuburan," *J. Tek. Inform. Stmik Antar Bangsa*, vol. 5, no. 1, pp. 51–59, 2019.
[22] U. K. Singh and M. Rout, "Genetic Algorithm based Feature Selection to Enhance Breast Cancer Classification," *Proc. IEEE InC4 2023 - 2023 IEEE Int. Conf. Contemp. Comput. Commun.*, vol. 1, pp. 1–5, 2023, doi: 10.1109/InC457730.2023.10263100.
[23] A. R. Naufal and A. T. Suseno, "Penerapan Fitur Seleksi dan Particle Swarm Optimization pada Algoritma Support Vector Machine untuk Analisis Credit Scoring," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 184–195, 2023, doi: 10.47065/josyc.v5i1.4409.
[24] S. I. Novichasari and I. S. Wibisono, "Particle Swarm Optimization For Improved Accuracy of Disease Diagnosis," *J. Appl. Intell. Syst.*, vol. 5, no. 2, pp. 57–68, 2021, doi: 10.33633/jais.v5i2.4242.
[25] D. Saputra, W. S. Dharmawan, and W. Irmayani, "Performance Comparison of the SVM and SVM-PSO Algorithms for Heart Disease Prediction," *Int. J. Adv. Data Inf. Syst.*, vol. 3, no. 2, pp. 74–86, 2022, doi: 10.25008/ijadis.v3i2.1243.
[26] W. Yuliani and E. Supriatna, *Metode Penelitian Bagi Pemula*. Penerbit Widina Bhakti Persada Bandung, 2023.
[27] Rukminingsih, G. Adnan, and M. Adnan Latief, *Metode Penelitian Pendidikan (Kuantitatif, Kualitatif & Penelitian Tindakan Kelas)*. Yogyakarta: Erhaka Utama, 2020.
[28] T. A. Y. Siswa, *Data Mining: Mengupas Tuntas Analisis Data Dengan Metode Klasifikasi Hingga Deployment Aplikasi Menggunakan Python*. Samarinda: UMKT PRESS, Universitas Muhammadiyah Kalimantan Timur, 2023.
[29] T. T. Wong and P. Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, 2020, doi: 10.1109/TKDE.2019.2912815.
[30] I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, "Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem," *ACM Int. Conf. Proceeding Ser.*, pp. 412–419, 2021, doi: 10.1145/3453892.3461323.
[31] I. R. Pratama, M. Maimunah, and E. R. Arumi, "Sistem Klasifikasi Penjualan Produk Alat Listrik Terlaris Untuk

Optimasi Pengadaan Stok Menggunakan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 6, no. 4, p. 2135, 2022, doi: 10.30865/mib.v6i4.4418.

[32] A. Arora *et al.*, "Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India," *Sci. Total Environ.*, vol. 750, p. 141565, 2021, doi: 10.1016/j.scitotenv.2020.141565.

[33] G. Eom and H. Byeon, "Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique," *Mathematics*, vol. 11, no. 16, p. 3605, 2023, doi: 10.3390/math11163605.

[34] P. Wibowo and C. Fatichah, "An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 7, no. 1, pp. 63–71, 2021, doi: 10.26594/register.v7i1.2206.

[35] S. Kanwal, J. Rashid, M. W. Nisar, J. Kim, and A. Hussain, "An Effective Classification Algorithm for Heart Disease Prediction with Genetic Algorithm for Feature Selection," *Proc. 2021 Mohammad Ali Jinnah Univ. Int. Conf. Comput. MAJICC 2021*, no. April, 2021, doi: 10.1109/MAJICC53071.2021.9526242.

[36] F. Faldi, T. NurHalisha, W. J. Pranoto, and ..., "The application of particle swarm optimization (PSO) to improve the accuracy of the naive bayes algorithm in predicting floods in the city of Samarinda," *J. Intell. ...*, vol. 6, no. 3, pp. 138–146, 2023, [Online]. Available: http://idss.iocspublisher.org/index.php/jidss/article/view/148%0Ahttps://idss.iocspublisher.org/index.php/jidss/article/download/148/99

[37] F. Maulidina, Z. Rustam, and J. Pandelaki, "Lung Cancer Classification using Support Vector Machine and Hybrid Particle Swarm Optimization-Genetic Algorithm," *2021 Int. Conf. Decis. Aid Sci. Appl. DASA 2021*, pp. 751–755, 2021, doi: 10.1109/DASA53625.2021.9682259.