Vol. 5 No. 1, 2024, pp. 27-37 **Jurnal Pepadun**

Implementasi SMOTE dan *Support Vector Machine* Pada Klasifikasi Data Tidak Seimbang Metilasi Arginin

[*1]Favorisen R. Lumbanraja, [2]Ester Caroline Lumban Gaol, [3]Dewi Asiah Shofiana dan [4]Akmal Junaidi

[1,2,3,4]Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung
Jl. Prof. Sumantri Brojonegoro No. 1, Kel. Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung, Indonesia, 35141

e-mail: [*1]favorisen.lumbanraja@fmipa.unila.ac.id, [2]esterclumbangaol@gmail.com [3]dewi.asiah@fmipa.unila.ac.id, [4]akmal.junaidi@fmipa.unila.ac.id

Abstract — *Imbalanced data is one of the crucial problems in machine learning and data mining which may provide low accuracy in minority classes and makes the classification method not fully optimized. The Arginine Methylation dataset for example, gives a large amount of imbalanced data. Methylation is one of the post-translational modification processes that occurs in arginine protein which affects signal transduction and RNA binding inside cytoplasms. Therefore, it is essential to handle imbalanced data for classification. Synthetic Minority Oversampling Technique (SMOTE) is an algorithm for solving imbalanced data in classification using the concept of k-nearest neighbors. Support Vector Machine (SVM) is a supervised learning method which splits datasets using hyperplane and maximize margin distance. In this research, the arginine methylation dataset is divided into three experimental data, which consists of training data, testing data, and independen data. Data processing goes through a series of steps; data pre-processing (clean redundance data), feature extraction (generates 159 feature dimensions), SMOTE and SVM modeling, and classification testing using 10-fold cross-validation and confusion matrix. The accuracy of training data is 100% in RBF kernel, whereas testing data gives a low accuracy of 65,90% in linear kernel. Independen data have decent accuracy in linear kernel by 98,50% percentage.*

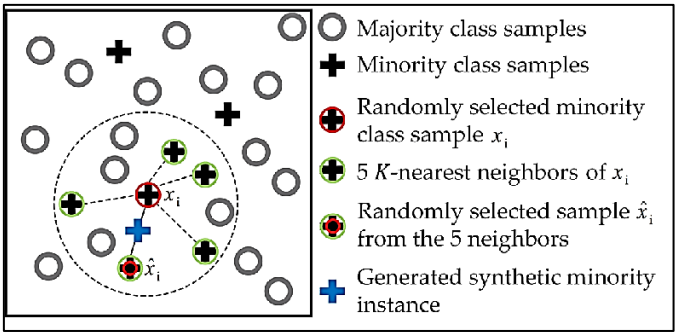
Keywords: *Imbalanced Data*; *Methylation*; *Post-Translational Modification*; *SMOTE*; *Support Vector Machine*.

1. PENDAHULUAN

Modifikasi pasca-translasi atau *post-translational modifications* (PTM) merupakan modifikasi yang terjadi akibat perubahan rantai asam amino pada proses translasi sintesis protein [1]. PTM dapat mempengaruhi fungsi protein baik aktivitas, lokalisasi hingga interaksi protein [2]. Urutan asam amino menentukan proses identifikasi dan penggambaran proses PTM seperti: fosforilasi, asetilasi, glikosilasi, metilasi, ubiquitinasi, nitrosilasi, dan lipidasi [3]. Metilasi Arginin merupakan jenis PTM yang umum terjadi pada tubuh manusia yang menghasilkan guanidina (basa organik) yang berikatan antara gugus metil dan atom nitrogen Arginin [4]. Riset mengenai Metilasi Arginin telah banyak dikembangkan, beberapa diantaranya adalah riset Kumar et al. (2017) [5] dan Lumbanraja et al. (2019) [6]. Riset Kumar et al. (2017) [5] yang berjudul "*PRmePRed: A Protein Arginine Methylation Prediction Tool*" menganalisis metilasi arginin menggunakan metode *Support Vector Machine*. Data protein Arginin diperoleh dari basis data UniProt (*release* 2015_06) yang terdiri data data uji, data latih, dan data independen. Adapun terdapat ketidakseimbangan data yang digunakan terlihat pada data latih dengan perbandingan positif:negatif sebesar 1:4. Hasil penelitian menunjukkan tingkat akurasi sebesar 84,10% untuk data uji, 90% untuk data latih, dan 93% untuk data independen. Hasil sensitivitas sebesar 82,38%, spesifisitas 83,77%, dan MCC 66,20%.

Penelitian selanjutnya oleh Lumbanraja et al. (2019) [6] merupakan pengembangan riset dari penelitian Kumar et al., (2017) [5] dengan membandingkan kinerja metode SVM dengan metode *random forest* serta menggunakan *feature extraction* yang berbeda. Hasil akurasi yang diperoleh untuk data latih sebesar 93,76%, data uji 80,32%, dan data independen 98,08%. Pada data independen, data positif dan negatif yang digunakan memiliki perbandingan yang cukup besar yakni 8:92. Oleh karena itu, akurasi pada data independen dinilai kurang baik karena lebih banyak menggunakan data negatif daripada data positif. Hal ini mengindikasikan terjadinya ketidakseimbangan data pada dua penelitian sebelumnya. *Imbalanced data* merupakan

ketidakseimbangan jumlah pada kelas data tertentu yang menunjukkan adanya kelas minoritas dan kelas mayoritas [7]. Penggunaan data tidak seimbang pada proses klasifikasi, akan berdampak pada hasil akurasi yang lebih rendah pada kelas minoritas[8], dan kualitas metode klasifikasi yang buruk. Oleh karena itu, ketidakseimbangan data perlu ditangani untuk memperbaiki hasil klasifikasi. Pada dua penelitian sebelumnya, penanganan ketidakseimbangan data adalah melakukan *undersampling data*, yaitu pengurangan jumlah data kelas mayoritas agar memiliki jumlah yang setara dengan kelas minoritas. Pengurangan data pada teknik *undersampling* dapat menghilangkan sebagian informasi penting pada data dan memungkinkan terjadinya *underfitting*.



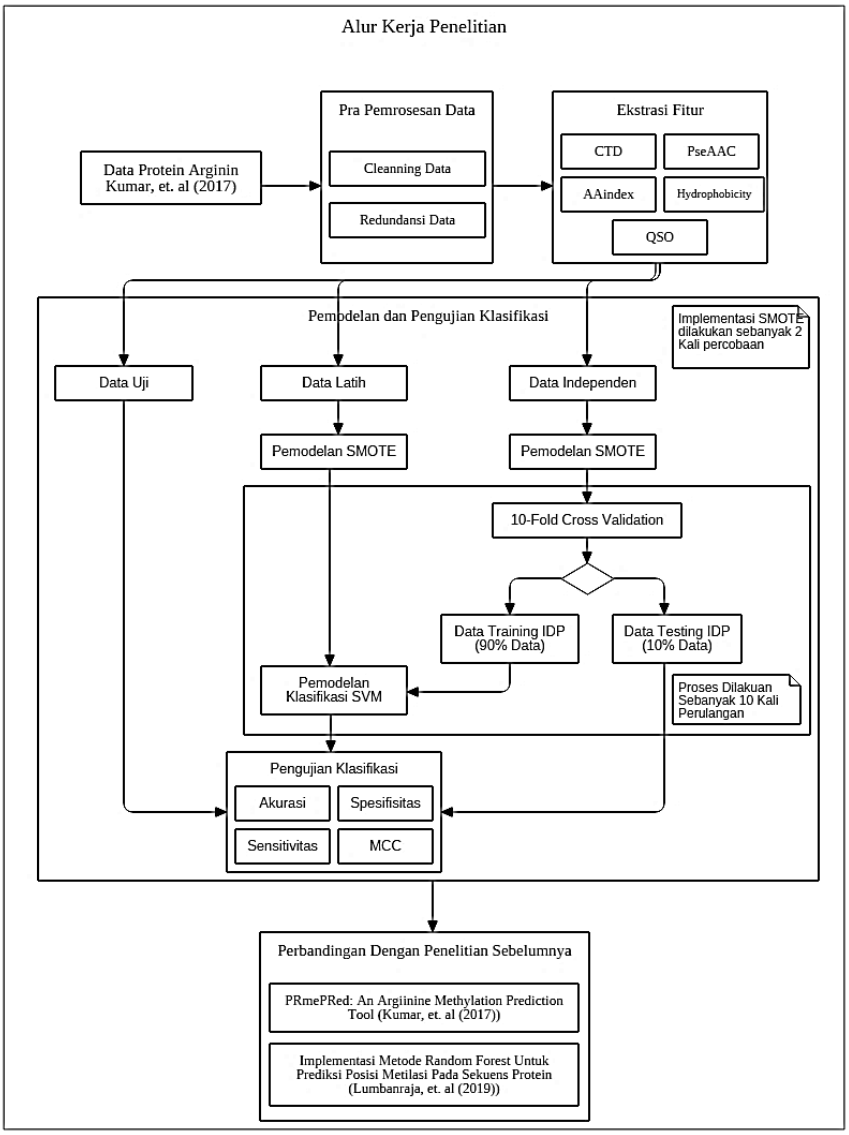
Gambar 1. Konsep Algoritme SMOTE [8].

Teknik penyelesaian lain untuk kasus data tidak seimbang adalah menggunakan algoritme pendukung selain metode klasifikasi *machine learning*. Metode *Synthetic Minority Oversampling Technique* (SMOTE) merupakan metode *oversampling* yang bekerja membuat salinan data minoritas (disebut data sintetik) dengan memanfaatkan konsep algoritme *k-nearest neighbor*. SMOTE dapat memperbaiki performa metode klasifikasi dan menghasilkan tingkat akurasi yang lebih baik [9]. SMOTE dapat diuji pada data yang memiliki tingkat ketidakseimbangan dan jumlah data yang berbeda [10]. *Oversampling* merupakan peningkatan pada data kelas minoritas sehingga memiliki jumlah mendekati kelas mayoritas. Hal ini berarti tidak ada pengurangan jumlah data baik pada kelas mayoritas ataupun minoristas sehingga tidak ada informasi yang hilang.

Dalam penelitian ini, metode klasifikasi yang digunakan adalah *support vector machine* yang memisahkan data menggunakan *hyperplane* (garis pemisah) dan mengoptimalkan *margin* (jarak pemisah) antara dua kelas. SVM digunakan baik untuk klasifikasi data dan keperluan regresi yang dapat membagi data linear maupun non-linear secara baik dalam aplikasi ilmiah dan teknik [11]. Oleh karena itu, penelitian ini berfokus meningkatan kelas minoritas pada data metilasi protein arginin menggunakan algoritme SMOTE dan melakukan klasifikasi kembali menggunakan metode *Support Vector Machine*.

2. METODOLOGI PENELITIAN

Alur pengerjaan penelitian klasifikasi *imbalanced data* menggunakan algoritme SMOTE dan metode *Support Vector Machine* (SVM) pada kasus metilasi *sequence* protein Arginin divisualisasikan pada Gambar 2.



Gambar 2. Alur Pengerjaan Penelitian.

2.1. Pengumpulan Data

Data penelitian didapat dari riset Kumar et al. (2017) [5] dari *database* UniProt (*release* 2016_05). Data metilasi protein Arginin tersebut terdiri dari 6.581 data negatif dan 4.331 data positif dengan total 10.912 data dan panjang *sequence* 19. Data terbagi menjadi data latih, data uji, dan data independen yang terinci pada Tabel 1.

Tabel 1. Data Metilasi Protein Arginin [5].

Tipe	Jenis Data	Jumlah *Sequence* Arginin
Data Latih	Positif	1.038
	Negatif	5.190
Data Uji	Positif	260
	Negatif	260
Data Independen	Positif	1.131
	Negatif	3.033

2.2. Prapemrosesan Data

Preprocessing terbagi menjadi dua tahap, yaitu *cleaning* dan redundansi data. Tahap *cleaning data* dilakukan dengan menghapus huruf X yang tidak termasuk dalam daftar nama asam amino pada sekuens protein. Selanjutnya, data yang sudah dibersihkan dilakukan redundansi dengan CD-HIT sebesar 40% *cut-off*. Redundansi bertujuan untuk menghilangkan sebagian sekuens yang memiliki kemiripan dengan sekuens lain untuk menghindari tingginya daya komputasi. Nilai *threshold* sebesar 40% *cut-off* merujuk pada sekuens yang memiliki kemiripan di atas 40% akan dihapus dari urutan.

2.3. Ekstraksi Fitur

Ekstraksi fitur merupakan tahap mengekstrak sekuens protein yang berupa data *string* menjadi data numerik, agar dapat diproses pada model klasifikasi [12]. Fitur mengekstraksi asam amino dengan mempertimbangkan probabilitas urutran asam amino [13]. Sekuens protein diekstrak berdasarkan sifat fisikokimia dan biokimia yang terdiri dari:

a. *Composition, Transition, and Distribution:* ekstrak berdasarkan sifat fisikokimia dan biokimia protein.
b. AAindex: ekstrak fisikokimia dan biokimia dari basis data AAindex.
c. *Hydrophobicity*: mengukur daya tarik protein terhadap lapisan lipid (hidropobisitas).
d. *Pseudo Amino Acid Composition*: prediksi lokalisasi subseluler dan tipe membran protein.
e. *Quasi-Sequence Order*: prediksi lokalisasi subseluler berdasarkan sifat fisikokimia dengan pendekatan statistik.

2.4. Pemodelan SMOTE dan SVM

Pemodelan SMOTE dilakukan pada data latih dan data independen sebanyak dua kali. Pada data uji tidak dilakukan SMOTE, karena data uji menerapkan pemodelan SMOTE yang dilakukan pada data latih. Khusus pada data independen, setelah menerapkan SMOTE dilakukan pembagian data menjadi *training data* dan *testing data* menggunakan skema *K-fold cross-validation* dengan nilai $K = 10$. *Cross-validation* merupakan evaluasi kinerja metode *supervised learning* yang memperkirakan *test error rate* dan *training error rate* [14]. *K-fold cross-validation* membagi data menjadi sejumlah k *fold* yang sama, dimana *fold* pertama sebagai data uji, sedangkan k-1 *fold* yang lain sebagai data latih. Nilai k=10 pada penelitian ini menunjukkan $\frac{1}{10}$ *fold* adalah data uji, sedangkan $\frac{9}{10}$ *fold* lainnya adalah data latih. Data latih dan data independen metilasi arginin yang telah di*oversampling*, selanjutnya diklasifikasi menggunakan metode *Support Vector Machine*. Klasifikasi menggunakan empat *kernel* SVM yakni: linear, *polynomial, radial basis function,* dan *sigmoid.* Uraian mengenai keempat *kernel* SVM tersebut disajikan dalam Tabel 2.

Tabel 2. *Kernel Support Vector Machine.*

Kernel	**Rumus**
Linear	$K(x_i.x_j) = x_i.x_j$
Polynomial	$K(x_i.x_j) = (\gamma(x_i.x_j) + r)^d, \gamma \geq 0$
Radial Basis Function	$K(x_i.x_j) = \exp(-\gamma\left\|x_i - x_j\right\|^2), \lambda \geq 0$
Sigmoid	$K(x_i.x_j) = \tanh(\gamma(x_i..x_j) + r), \gamma \geq 0$

2.5. Pengujian Klasifikasi

Data metilasi arginin yang telah diklasifikasi dilakukan pengujian menggunakan *confusion matrix. Confusion matrix* merupakan teknik evaluasi klasifikasi yang menunjukkan hasil prediksi secara aktual dari performa prediktor [12]. *Confusion matrix* untuk dua kelas data dapat dilihat pada Tabel 3.

Tabel 3. *Confusion Matrix* [12].

Predicted	***Actual***	
	Event	***No-Event***
Event	TP	FP
Non-event	FN	TN

Parameter pengujian klasifikasi menggunakan empat indikator, yaitu: akurasi (ACC), sensitivitas (SN), spesifisitas (SP), dan *Matthew Correlation Coeficient* (MCC). Akurasi merupakan kedekatan nilai prediksi dengan nilai sebenarnya, sensitivitas adalah persentase data positif yang terklasifikasikan dengan benar dari jumlah keseluruhan data positif. Spesifisitas menunjukkan persentase data negatif benar dari keseluruhan data negatif, sedangkan MCC menunjukkan nilai kualitas klasifikasi sebuah prediktor [15]. Berikut adalah rumus perhitungan untuk keempat indikator pengukuran pada Persamaan 1 hingga Persamaan 4.

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} x100\% \quad (1)$$

$$SN = \frac{TP}{TP + FN} x100\% \quad (2)$$

$$SP = \frac{TN}{TN + FP} x100\% \quad (3)$$

$$MCC = \frac{TPxFN - FPxFN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1. Prapemrosesan Data

Prapemrosesan data dilakukan dengan membersihkan data X pada sekuens protein dan meredundansi menggunakan CD-HIT sebesar 40% *cut-off*. Berikut adalah hasil prapemrosesan data metilasi arginin pada setiap tahap prapemrosesan yang dimuat dalam Tabel 4.

Tabel 4. Data Metilasi Protein Arginin Sebelum dan Setelah Prapemrosesan Data.

Data Metilasi Protein Arginin		**Data Awal**	***Cleanning***	**Redundansi**
Latih	Positif	1.038	1.018	1.038
	Negatif	5.190	5.135	5.055
Uji	Positif	260	253	253
	Negatif	260	257	257
Independen	Positif	1.131	1.006	486
	Negatif	3.033	3.030	1.721

Tabel 4 menunjukkan pengurangan pada setiap prapemrosesan data. Pada *cleaning data* terjadi penurunan 1,9% dari 10.912 menjadi 10.699. Selanjutnya, pada tahap redundansi data terjadi penururunan signifikan sebesar 17,84% menjadi 8.790.

3.2. Ekstraksi Fitur

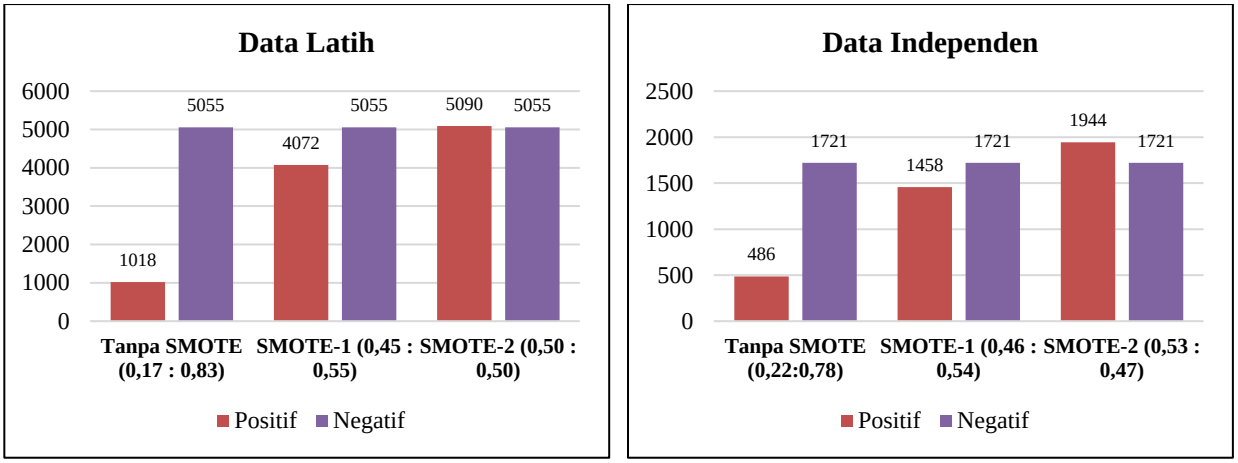
Ekstraksi *string* sekuens protein menggunakan 6 fitur fisikokimia dan biokimia yang masing-masing menghasilkan jumlah dimensi berbeda yang terangkum dalam Tabel 5. Dimensi tersebut memuat nilai ekstraksi asam amino dengan rentang nilai 0-1. Nilai 0 merujuk pada asam amino yang tidak mendekati sifat biokimia atau fisikokimia tertentu, sedangkan nilai 1 mendekati sifat tersebut.

Tabel 5. Dimensi Ekstraksi Fitur.

Fitur Ekstraksi	Dimensi
CTD	21
AAindex	19
Hydrophobicity	19
PseAAC	24
QSO	76
Jumlah	**159**

3.3. Pemodelan SMOTE

Peningkatan data sintetik pada kelas minoritas dilakukan sebanyak dua kali pada data latih dan data independen. Hal ini dilakukan untuk melihat performa *classifier* yang digunakan ketika data kelas minoritas memiliki jumlah yang mendekati dan melebihi kelas mayoritas.



(a) (b)

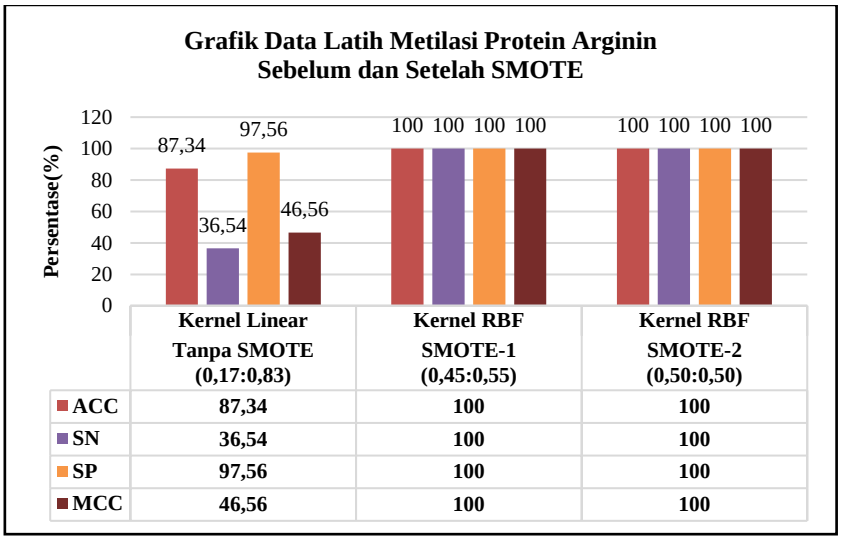
Gambar 3. (a) Pemodelan SMOTE Data Latih dan (b) Data Independen.

Gambar 3 memvisualisasikan peningkatan data minoritas pada data latih dan data independen. Pada Gambar 3(a) menunjukkan kenaikan data minoritas dengan persentase 75% pada pemodelan SMOTE pertama (selanjutnya SMOTE-1), sehingga perbandingan data positif:negatif menjadi 0,45:0,55 dari 0,17:0,83. Pemodelan SMOTE kedua (selanjutnya SMOTE-2) berhasil meningkatkan 80% data sintetik dengan hasil perbandingan data menjadi 0,50:0,50. Untuk data independen yang dilihat pada Gambar 3(b), pemodelan SMOTE pertama ditingkatkan sebesar 66,7% dengan rasio akhir menjadi 0,46:0,54. Pemodelan SMOTE kedua data independen menyintesiskan 75% data minoritas dengan hasil perbandingan akhir menjadi 0,53:0,47.

3.4. Hasil Klasifikasi

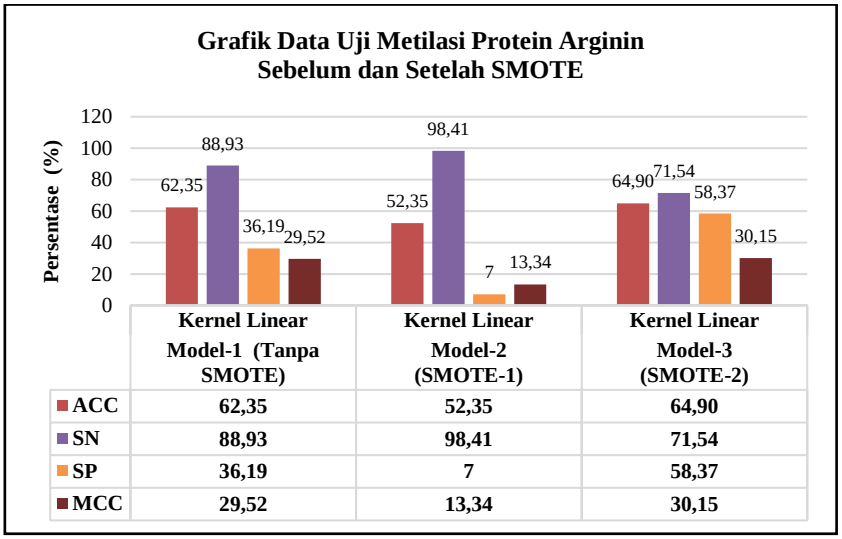
Data metilasi protein arginin diklasifikasi menggunakan metode *support vector machine* terhadap empat *kernel*, yakni: linear, *polynomial*, *radial basis function*, dan *sigmoid*. Setiap model SMOTE memberikan nilai pengujian yang beragam tergantung *kernel* yang digunakan. Pada penelitian berikut, diberikan gambaran hasil pengujian terbaik pada *kernel* yang digunakan pada model SMOTE.

Berdasarkan Gambar 4, data latih menunjukkan peningkatan signifikan pada setiap indikator pengujian setelah penerapan SMOTE. Baik pada model SMOTE-1 dan SMOTE-2 memiliki rata-rata nilai pengujian 100% pada masing-masing akurasi, sensitivitas, spesifisitas, dan MCC. Berikutnya adalah grafik hasil pengujian data uji yang menerapkan pemodelan SMOTE dari data latih yang disajikan dalam Gambar 5.



	Kernel Linear Tanpa SMOTE (0,17:0,83)	Kernel RBF SMOTE-1 (0,45:0,55)	Kernel RBF SMOTE-2 (0,50:0,50)
ACC	87,34	100	100
SN	36,54	100	100
SP	97,56	100	100
MCC	46,56	100	100

Gambar 4. Grafik Pengujian Klasifikasi Data Latih.

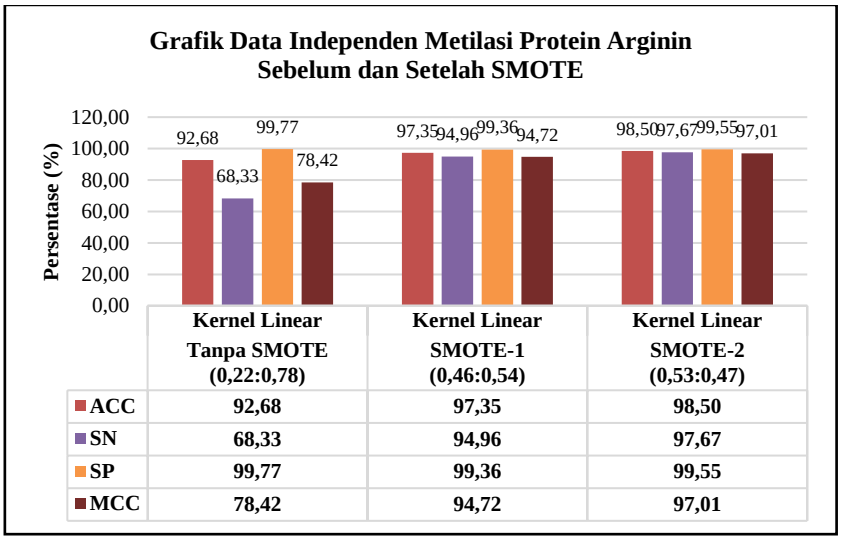


	Kernel Linear Model-1 (Tanpa SMOTE)	Kernel Linear Model-2 (SMOTE-1)	Kernel Linear Model-3 (SMOTE-2)
ACC	62,35	52,35	64,90
SN	88,93	98,41	71,54
SP	36,19	7	58,37
MCC	29,52	13,34	30,15

Gambar 5. Grafik Pengujian Klasifikasi Data Uji.

Visualisasi Gambar 5 memperlihatkan nilai pengukuran yang diperoleh data uji tidak sebaik data latih pada Gambar 4. Pada Model-2 (penerapan SMOTE pertama) terjadi penurunan nilai setiap indikator dari Model-1 (tanpa SMOTE). Selanjutnya, pada Model-3 (penerapan SMOTE kedua) menujukkan kenaikan tidak terlalu tinggi jika dibandingkan dengan hasil pengukuran Model-1. Hal ini ditunjukkan dengan rata-rata nilai akurasi tertinggi data uji adalah 64,90% dibandingkan data latih yang memperoleh nilai 100%. Selain itu, nilai MCC pada Model-2 dan Model-3 yang rendah menujukkan lemahnya kinerja prediktor pada data uji. Masalah mengenai hasil data uji lebih rendah daripada data latih yang memiliki selisih terlampau jauh, mengindikasikan terjadinya kesalahan *overfitting.*

Overfitting merupakan keadaan model klasifikasi yang bekerja baik (*fit*) hanya pada data latih, sedangkan ketika diterapkan pada data uji akan memberikan kinerja yang tidak sebanding dengan data latih. Hal seperti ini terjadi karena kemungkinan banyaknya jumlah dimensi yang digunakan pada saat proses klasifikasi yang mengakibatkan jarak antar data semakin jauh dikenal dengan istilah *curse of dimensional. Overfitting* dapat diatasi dengan menyederahanakan model klasifikasi dan melakukan seleksi fitur pada objek data. Terakhir visualisasi pengujian data independen pada penerapan SMOTE yang menggunakan skema 10-*fold cross-validation* dalam pembagian *training data* dan *testing data* dalam Gambar 6.



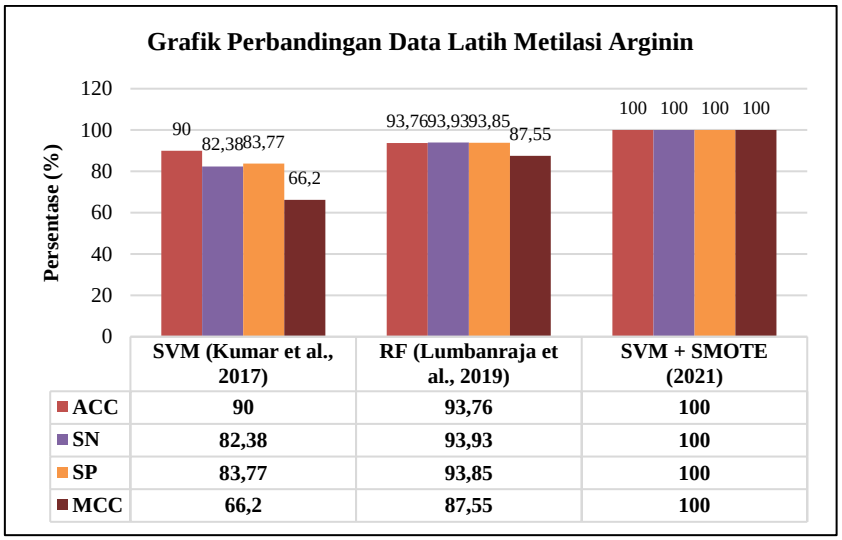
	Kernel Linear Tanpa SMOTE (0,22:0,78)	Kernel Linear SMOTE-1 (0,46:0,54)	Kernel Linear SMOTE-2 (0,53:0,47)
ACC	92,68	97,35	98,50
SN	68,33	94,96	97,67
SP	99,77	99,36	99,55
MCC	78,42	94,72	97,01

Gambar 6. Grafik Pengujian Klasifikasi Data Independen.

Berdasarkan Gambar 6, memperlihatkan peningkatan hasil pengujian data independen setelah penerapan SMOTE. Hasil pengujian terbaik didapat setelah penerapan SMOTE kedua kali dengan rata-rata nilai akurasi 98,50%, sensitivitas 97,67%, spesifisitas 99,55%, dan MCC 97,01%.

3.5. Perbandingan Penelitian Sebelumnya

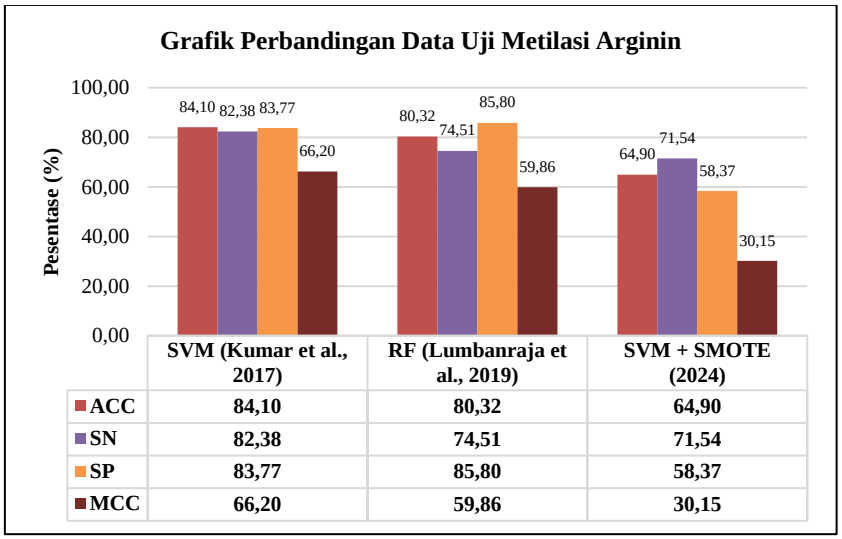
Hasil yang didapat pada penelitian ini, selanjutnya dilakukan perbandingan dengan dua penelitian sebelumnya guna mendapatkan kesimpulan terhadap penggunaan algoritme SMOTE dalam mengatasi ketidakseimbangan data. Dua penelitian tersebut adalah penelitian Kumar et al. (2017) [5] dan Lumbanraja et al. (2019) [6] yang sekaligus menjadi referensi dalam penelitian ini. Berikut disajikan grafik perbandingan hasil pengujian data metilasi protein arginin pada empat indikator pengukuran yang dimuat dalam Gambar 7 hingga Gambar 9.



	SVM (Kumar et al., 2017)	RF (Lumbanraja et al., 2019)	SVM + SMOTE (2021)
ACC	90	93,76	100
SN	82,38	93,93	100
SP	83,77	93,85	100
MCC	66,2	87,55	100

Gambar 7. Grafik Perbandingan Pengujian Data Latih.

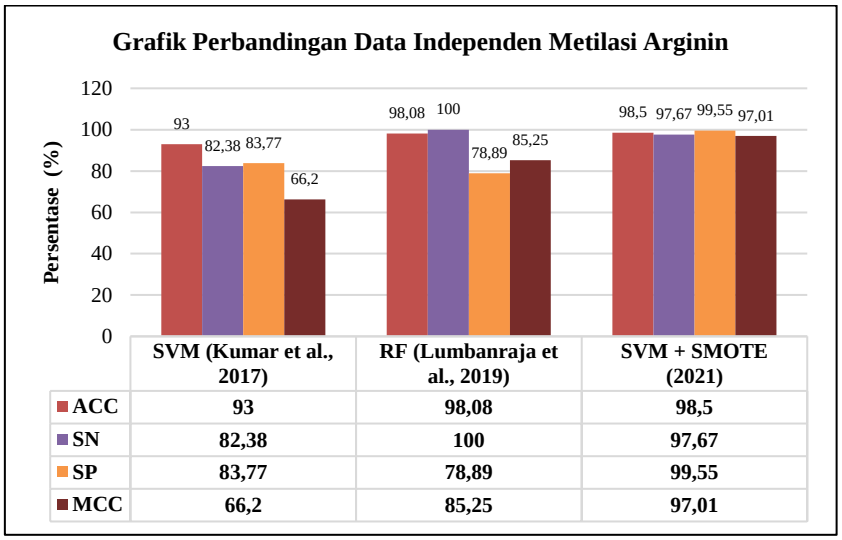
Gambar 7 menjelaskan bahwa penerapan SMOTE bekerja dengan baik pada data latih metilasi protein arginin. Hal ini dapat dilihat dari nilai setiap indikator lebih baik dari dua penelitian sebelumnya. Nilai akurasi, sensitivitas, spesifisitas dan MCC berturut-turut adalah 100%.



	SVM (Kumar et al., 2017)	RF (Lumbanraja et al., 2019)	SVM + SMOTE (2024)
ACC	84,10	80,32	64,90
SN	82,38	74,51	71,54
SP	83,77	85,80	58,37
MCC	66,20	59,86	30,15

Gambar 8. Grafik Perbandingan Pengujian Data Uji.

Hasil pengujian data uji penelitian ini memperoleh hasil terendah pada setiap indikator pengukuran. Terutama pada spesifisitas yang menjadi fokus penelitian ini, masih memberikan nilai rendah dibandingkan dengan dua penelitian sejenisnya. Akurasi yang diberikan menunjukkan angka 64,90%, denagn tingkat sensitivitas 71,54%, spesifisitas 58,37%, dan MCC 30,15%. Hal ini terjadi akibat *overfitting* antara data latih dan data uji. Nilai pengukuran terbaik masih diperoleh pada riset Kumar et al. (2017) [5] dengan nilai akurasi rata-rata 84,10%, sensitivitas 82,38%, spesifisitas 84,77%, dan MCC 66,20%.



	SVM (Kumar et al., 2017)	RF (Lumbanraja et al., 2019)	SVM + SMOTE (2021)
ACC	93	98,08	98,5
SN	82,38	100	97,67
SP	83,77	78,89	99,55
MCC	66,2	85,25	97,01

Gambar 9. Grafik Perbandingan Pengujian Data Independen.

Berdasarkan Gambar 9, perbandingan data independen metilasi protein arginin terhadap dua penelitian sebelumnya mendapat hasil terbaik pada penelitian ini. Pada indikator akurasi, spesifisitas, dan MCC memperoleh rata-rata tertinggi dengan nilai 98,5%, 97,67%, 99,55%, dan 97,01% berturut-turut. Untuk nilai sensitivitas tertinggi masih diperoleh pada penelitian Lumbanraja et al. (2019) [6] dengan nilai 100%.

4. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan pada data tidak seimbang metilasi protein arginin menggunakan algoritme SMOTE dan *Support Vector Machine*, dapat disimpulkan bahwa terjadi *overfitting* pada data latih dan dan data uji. Hal ini ditunjukkan pada nilai akurasi, sensitivitas, spesifisitas, dan MCC pada data latih adalah 100% berturut-turut. Namun, pada data uji, terjadi selisih cukup jauh dengan hasil akurasi 64,9%, sensitivitas, 71,54%, spesifisitas 58,37%, dan MCC 30,15%. Hal ini terjadi akibat banyaknya jumlah dimensi yang digunakan pada proses klasifikasi sehingga mengakibatkan kinerja *classifier* menjadi tidak efektif (*curse of dimensional*). Pada data independen, algoritme SMOTE dan SVM bekerja dengan optimal yang dibuktikan dengan akurasi sebesar 98,5%, sensitivitas 97,675, spesifisitas 99,55%, dan MCC 97,01%. Adapun saran untuk penelitian sebelumnya adalah dapat mengimplementasikan teknik seleksi fitur untuk mengurangi kesalahan *overfitting* yang dapat diimplementasikan pada algoritme penyelesaian ketidakseimbangan data dan metode klasifikasi lainnya.

DAFTAR PUSTAKA

[1] Q. Li & S. Shah, Structure-Based Virtual Screening BT - Protein Bioinformatics: From *Protein Modifications and Networks to Proteomics*, 2017.

[2] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, & E. Skrzypek, PhosphoSitePlus, 2014: Mutations, PTMs & Recalibrations, *Nucleic Acids Res*earch, Vol. 43, no. D1, pp. D512–D520, 2015, doi: 10.1093/nar/gku1267.

[3] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, & K. C. Chou, iPTM-mLys: Identifying Multiple Lysine PTM Sites and Their Different Types, *Bioinformatics*, Vol. 32, No. 20, pp. 3116–3123, 2016, doi: 10.1093/bioinformatics/btw380.

[4] J. D. Gary & S. Clarke, RNA and Protein Interactions Modulated by Protein Arginine Methylation., *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 61. pp. 65–131, 1998, doi: 10.1016/s0079-6603(08)60825-9.

[5] P. Kumar, J. Joy, A. Pandey, & D. Gupta, PRmePRed: A protein Arginine Methylation Prediction Tool," *PLoS One*, Vol. 12, No. 8, 2017, doi: 10.1371/journal.pone.0183318.

[6] F. R. Lumbanraja, W. Mudyaningsih, B. Hermanto, & A. Syarif, Implementasi Metode Random Forest Untuk Prediksi Posisi Metilasi Pada Sekuens Protein, in *Seminar Nasional Sains, Matematika, Informatika, dan Aplikasinya*, 2019, pp. 105–112.

[7] M. R. Faisal, *Seri Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R*, no. February. 2016.

[8] N. Noorhalim, A. Ali, & S. M. Shamsuddin, Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE, in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, 2019, pp. 19–30.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[10] O. Maimon & L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, 2010.

[11] J. Terzic, E. Terzic, R. Nagarajah, & M. Alamgir, *Ultrasonic Fluid Quantity Measurement in Dynamic Vehicular Applications: A Support Vector Machine Approach*, Springer Cham, 2013.

[12] M. Kuhn & K. Johnson, *Applied Predictive Modeling*, Springer New York, 2013.

[13] N. Bharill, A. Tiwari, & A. Rawat, A Novel Technique of Feature Extraction with Dual Similarity Measures for Protein Sequence Classification, in *Procedia Computer Science*, 2015, Vol. 48, No. C, pp. 795–801, doi: 10.1016/j.procs.2015.04.217.

[14] G. James, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer New York, 2000.

[15] M. Bekkar, H. K. Djemaa, & T. A. Alitouche, Evaluation Measures for Models Assessment over Imbalanced Data Sets, *Journal of Information Engineering and Application*, Vol. 3, No. 10, pp. 27–38, 2013.