Optimasi Algoritma SVM dalam Memprediksi Struktur Sekunder Protein menggunakan Metode *GridSearch Validation*

Metode Penelitian

Data Set

Penelitian ini menggunakan data RS126 yang merupakan data pasangan struktur primer dan struktur sekunder protein. Data RS126 tersebut diperoleh dari Laman Komunitas Data Kaggle(https://www.kaggle.com/datasets/tamzidhasan/rs126data).

Data RS126 merupakan salah satu dataset tertua dengan sejarah terpanjang untuk mengevaluasi prediksi struktur sekunder protein(Fai Chin et al., 2012). Dataset RS126 ini ditemukan pertama kali oleh Rost dan Sander pada tahun 1993 dan dipublikasikan dalam artikel ilmiah berjudul *Prediction of protein secondary structure at better than 70% accuracy* (Rost & Sander, 1993). Dataset RS126 terdiri dari 126 data protein yang memiliki panjang urutan protein rata-rata 185 asam amino. Dengan 32% dari RS126 adalah *alpha helix*, 21% adalah *beta*, dan 47% adalah *coil* (Tasari, Dinata Tarigan, et al., 2022).

Sampel pasangan struktur primer dan sekunder protein pada data RS126 ditunjukkan oleh tabel sampel dataset RS126 berikut.

Tabel 1. Sampel Dataset RS126

Structure	***Sequence***
Primary	APAFSVSPASGASDGQSVS VSVAAAGETYYIAQCAPVGG QDACNPATATSFTTDASGA ASFSFTVRKSYAGQTPSGTP VGSVDCATDACNLGAGNSG LNLGHVALTFG
Secondary	CCEEEEECCCCCCCCCEEE EEEECCCCEEEEEEECEEC CEECCCCCCCCCEEECCCCC CCEEEEECCCEEEEECCCC CEEEEEECCCCCCEEEEEC CCCCCCCCCCCCC

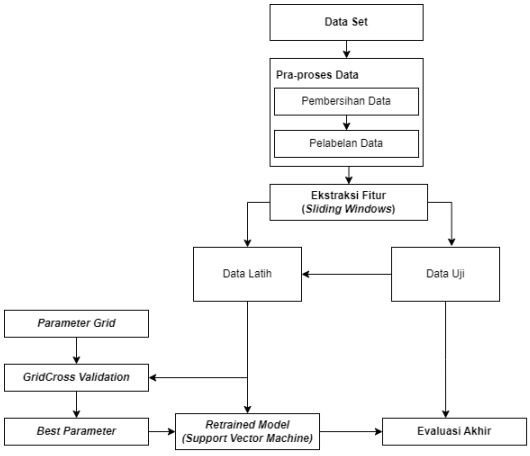
Tahapan Penelitian

Penelitian ini diawali dengan mempersiapkan data set yang akan digunakan, yakni dataset RS126. Setelah itu, dilanjutkan dengan *preprocessing* atau pra-prosess data, yang terdiri dari tahap pembersihan dan pelabelan pada data. Kemudian dilakukan ekstraksi fitur menggunakan teknik *sliding windows* lalu memisahkan data menjadi data latih dan data uji. Selanjutnya, menentukan nilai C, nilai *gamma*, dan kernel yang akan digunakan untuk menguji data pada model *support vector machine*. Adapun nilai parameter tersebut dapat dilihat pada tabel nilai *parameter grid* berikut.

Tabel 2. Nilai *parameter grid*

Parameter	Nilai
C	0.1, 1, 10, 100, 1000
Gamma (γ)	1, 0.1, 0.01, 0.001, 0.0001
Kernel	*Radial Basis Function*

Pada tahap berikutnya, data latih yang sudah disiapkan sebelumnya akan dibagi kembali menggunakan metode *Grid Cross Validation* untuk membuat model pasangan data yang lebih baik dengan melatih dan menguji semua bagian pada dataset pelatihan. Nilai *k-fold* yang digunakan pada teknik *Cross Validation* ini adalah k=5. Pada tahap ini juga akan dilakukan kombinasi antara data latih dengan nilai parameter yang sudah ditentukan untuk mencari nilai parameter terbaik. Setelah mendapatkan nilai parameter terbaik, model akan diuji kembali menggunakan data uji untuk selanjutnya dilakukan proses evaluasi menggunakan nilai *split test score, mean test score, rank test score*, dan akurasi



Gambar 1. Tahapan Penelitian

Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan salah satu metode yang sudah banyak diterapkan untuk berbagai jenis penelitian dibidang data dan text mining karena telah mampu menunjukkan perfoma yang lebih baik. Tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan pada svm sangat bergantung pada fungsi kernel dan parameter yang digunakan (Styawati & K,

2019). Berdasarkan fungsi kernelnya SVM dibagi menjadi:

a. Kernel Linear
   Fungsi kernel linear merupakan unutk memisahkan dua kelas pada hyperplane dengan soft margin (Purnami, Regresi, & Ordinal, 2012) Berikut ini merupakan persamaan kernel linear

$$K(x.x') = x.x'$$

b. Kernel Polynomial
   Kernel Polynomial merupakan kernel yang digunakan ketika data tidak terpisah secara linear (E, 2012) Berikut ini merupakan persamaan kernel polynomial

$$K(x_i.x_j) = (x_i.x_j) + c)^d$$

Kernel Radial Basis Function (RBF)

Metode klasifikasi SVM (Support Vector Machine) dengan RBF (Radial Basis Function) kernel merupakan metode yang sering digunakan karena memberikan hasil klasifikasi yang cukup akurat (Rarasmaya Indraswari, 2017).

Kernel RBF merupakan kernel yang digunakan Ketika data yang tidak dapat terpisah secara linier dimana dalam melakukan analisis RBF akan dilakukan optimasi parameter cost dan gamma (Widayani & Harliana, 2021). Berikut ini merupakan persamaan kernel RBF (Isnain, Sakti, Alita, & Marga, 2021):

$$K(x.x') = exp\ (\gamma \parallel x - x' \parallel^d)$$

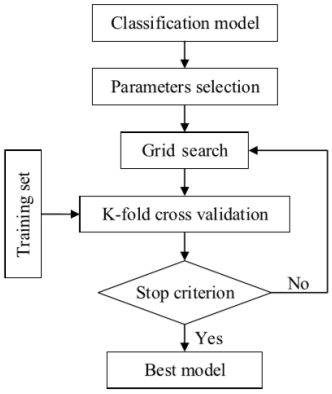
Hyperparameter Tuning

Hyperparameter merupakan variable yang mempengaruhi keluaran model. Hyperparameter tuning memiliki peran yang sangat penting dalam mengoptimalkan kinerja algoritma machine learning. Nilai hyperparameter tidak dapat ditentukan dari data yang selalu diberikan saat mendefenisikan model. Sebelum model menjalani proses, kita perlu menentukan nilai-nilai hyperparameter. Penelitian ini menggunakan teknik Grid Search untuk mencari nilai hyperparameter yang ideal pada model. Dimana kombinasi model dan hyperparameter diuji satu per satu dan dipilih dengan memvalidasi setiap kombinasi(Prima et al., 2021).

Grid Search

Grid Search merupakan salah satu algoritma yang biasa digunakan dalam meoptimisasi parameter. Cara kerja algoritma Grid Search ialah dengan membagi range parameter yang akan dioptimasi ke dalam grid dan melintasi semua titik untuk mendapatkan parameter yang optimal. Dalam penerapannya algoritma Grid Search harus dipandu oleh sejumlah matriks kinerja, biasanya diukur dengan cross validation pada data pelatihan(Yasin et al., 2016).

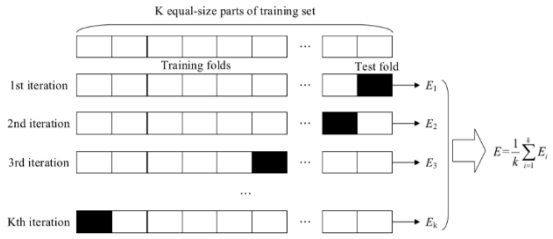
Kelebihan dari Algoritma Grid Search adalah dapat mencari parameter independen secara paralel, dan mengkonsumsi lebih sedikit waktu ketika parameter terbatas tersedia. Sementara kekurangan dari Algoritma Grid Search adalah perhitungan yang besar, terutama ketika lebih banyak parameter yang digunakan dan kesulitan dalam memperkirakan range parameter dan penalty factor (Wenwen et al., 2014)



Gambar 2. Proses Grid Search dan K-fold Cross Validation

K-fold Cross Validation

K-fold Cross Validation adalah metode yang umum digunakan untuk set training. Dimana metode ini bekerja dengan membagi set training menjadi K bagian yang berukuran sama. Setiap bagian dari set training dianggap sebagai set validation , dan bagian K−1 yang tersisa dianggap sebagai set train baru. Kemudian, model K akan dibentuk, dilatih, dan divalidasi dengan set training K−1 dan set validation K. Keakuratan model klasifikasi dengan semua kombinasi parameter lalu dibandingkan untuk menentukan kombinasi parameter pengklasifikasi yang optimal. Akhirnya, hyper-parameter disesuaikan dan dipasangkan dalam rentang parameter yang ditentukan pada data RS126, dan model klasifikasi dengan akurasi tertinggi dipilih oleh K-fold Cross Validation dan diterapkan dalam prediksi(Yan et al., 2022).

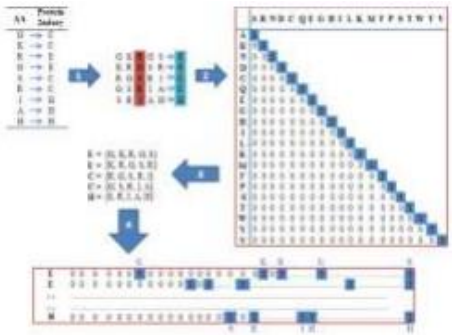


Gambar 3. Perhitungan K-fold Cross Validation

Sliding Window

Sliding Window merupakan sebuah ventilasi dengan ukuran piksel tertentu dalam lebar dan tinggi. Ukuran window itu sendiri bervariasi, tetapi proporsinya tetap sama. Window bergeser dari sudut kiri atas kearah kanan, lalu turun ke bawah dan bergerak ke kanan lagi dan juga ke sudut kanan bawah dan seterusnya yang menyerupai alphabet Z berulang angkah pertama pada ekstraksi fitur sliding window dimulai dengan pemilihan lebar window yaitu 15. Langkah kedua adalah pengambilan data dengan lebar window yang digunakan, setelah itu dilakukan ekstraksi pada setiap asam amino. Langkah terakhir adalah membangun inputan. Input yang digunakan untuk proses training adalah panjang fitur output ekstraksi asam amino dengan menggunakan lebar sliding window sebesar W*20, dimana 20 adalah ukuran pengkodean ortogonal berdasarkan struktur protein.(Tasari, Tarigan, et al., 2022).

Tahap sliding window juga membentuk segmentasi data yang kemudian dilakukan normalisasi data sebagai salah satu teknik preprocessing data sehingga semua data memiliki nilai(Sistem et al., 2021).



Gambar 4. Ilustrasi Sliding Window

Struktur Sekunder Protein

Ada empat jenis nukleotida (juga disebut basa). Mereka adalah Adenin (A). Timin (T), Sitosin (C) dan Guanin (G). Urutan nukleotida (juga disebut urutan DNA) tidak menentukan fungsi biologis sistem. Seperti disebutkan sebelumnya, fungsi ditentukan oleh struktur protein. Ada tiga tingkat struktur protein: struktur primer, sekunder dan tersier yang masing-masing satu, dua dan tiga dimensi. Struktur primer adalah urutan asam amino yang diperoleh dari urutan nukleotida.

Struktur sekunder memiliki tiga bentuk reguler: heliks alfa (α), lembaran beta (β) (kombinasi untaian beta) dan loop (juga disebut putaran balik atau gulungan). Dalam masalah prediksi struktur sekunder protein, inputnya adalah urutan asam amino sedangkan outputnya adalah struktur yang diprediksi (juga disebut konformasi, yang merupakan kombinasi dari heliks alfa, lembaran beta, dan loop). Sebuah protein khas mengandung sekitar 32% heliks alfa, 21% lembar beta dan 47% loop atau struktur non-reguler(Nanda et al., 2007).

Asam Amino

Asam amino memiliki dua gugus fungsi yaitu – NH2 dan – COOH Pada keadaan zwitter ion, biasanya gugus tersebut dalam keadaan – NH4 + dan – COO- . Kecuali prolin, 20 jenis asam amino pembentuk protein memiliki gugus karboksil bebas dan gugus amino bebas tidak tersubstitusi yang terikat pada atom karbon α sehingga dinamakan dengan α-asam amino. Berdasarkan strukturnya, 20 jenis asam amino pembentuk protein, 19 diantaranya merupakan amina primer dan 1 amina sekunder (prolin) (Kurniawati & Banowati, 2018).

Protein adalah rantai asam amino yang bergabung dengan ikatan peptida yang berperan penting dalam mengatasi berbagai masalah dalam tubuh manusia dan merupakan penyusun utama seluruh sel tubuh. Fungsi protein antara lain adalah membentuk enzim dan hormon, membentuk sel darah, dan membuat antibodi untuk melindungi tubuh dari penyakit dan infeksi(Wulandari, 2020)

Hasil dan Pembahasan

Penerapan Algoritma SVM

Pada penelitian sebelumnya (Tasari, Dinata Tarigan, et al., 2022), algoritma *support vector machine* mampu memprediksi struktur sekunder protein dengan baik pada nilai parameter C = 1 dan gamma(γ) = 0,1 dengan skor akurasi sebesar 0.62. Pada penelitian tersebut data dibagi menjadi lima rasio data berdasarkan ketentuan peneliti. Rasio data latih dan data uji tersebut dapat dilihat pada tabel 3 berikut.

Tabel 3. Pembagian rasio data latih & data uji

% Data Latih	% Data Uji	Kernel	C	γ	Nilai Akurasi
50	50	RBF	1	0,1	0.62

60	40	RBF	1	0,1	0.61
70	30	RBF	1	0,1	0.61
80	20	RBF	1	0,1	0.62
90	10	RBF	1	0,1	0.60

Sumber : Tasari, Dinata Tarigan, et al., (2022)

Sedangkan pada penelitian ini, algoritma *support vector machine* mampu memprediksi struktur sekunder protein dengan baik pada 3 kandidat parameter, yaitu C = 1000 dan gamma = 0.1, C = 100 dan gamma = 0.1}, serta C = 10 dan gamma = 0.1. Perolehan nilai rata-rata skor tes ketiga parameter tersebut dapat dilihat pada tabel berikut.

Tabel 4. Nilai rata-rata skor tes kandidat parameter

Parameter	Rata-rata skor tes
{'C': 1000, 'gamma': 0.1}	0.622553
{'C': 100, 'gamma': 0.1}	0.622553
{'C': 10, 'gamma': 0.1}	0.622553

Penelitian ini menggunakan nilai *sliding windows* sebesar 15 dan nilai *orthogonal encoding* sebesar 20 sehingga didapatkan nilai perataan matriks dua dimensi sebesar 300 yang akan digunakan untuk menyesuaikan ukuran label data dengan ukuran matriksasi data. Data yang sudah disesuaikan ukurannya kemudian dibagi menjadi data latih dan data uji dengan metode *K-fold Cross Validation*. Model SVM yang digunakan menggunakan kernel *Radial Basis Function* dengan lima nilai pada masing-masing parameter, yaitu parameter gamma = 1, 0.1, 0.01, 0.001, 0.0001 dan parameter C = 0.1, 1, 10, 100, 1000.

Optimalisasi Algoritma SVM

Penetapan lima nilai pada masing-masing parameter sebelumnya menghasilkan 25 kombinasi kandidat seperti pada tabel 5 berikut.

Tabel 5. Kandidat Parameter

Kandidat	Parameter	
	C	γ
Candidate1	0.1	1
Candidate2	0.1	0.1
Candidate3	0.1	0.01
Candidate4	0.1	0.001
Candidate5	0.1	0.0001
Candidate6	1	1
Candidate7	1	0.1
Candidate8	1	0.01
Candidate9	1	0.001
Candidate10	1	0.0001
Candidate11	10	1
Candidate12	10	0.1
Candidate13	10	0.01
Candidate14	10	0.001
Candidate15	10	0.0001
Candidate16	100	1
Candidate17	100	0.1
Candidate18	100	0.01
Candidate19	100	0.001
Candidate20	100	0.0001
Candidate21	1000	1
Candidate22	1000	0.1
Candidate23	1000	0.01
Candidate24	1000	0.001
Candidate25	1000	0.0001

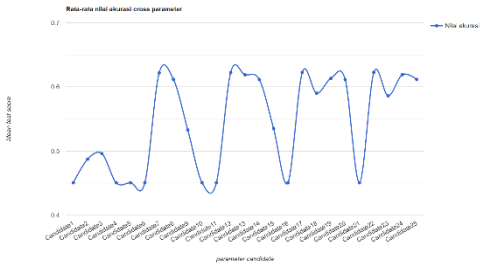
Masing-masing kandidat tersebut menghasilkan nilai rata-rata skor tes yang beragam, namun nilai rata-rata skor tes yang terbaik berada pada kandidat ke-22, ke-17, dan ke-12.

Tabel 6. Rata-rata skor tes

Kandidat	Rata-rata skor tes	Std	Rank
Candidate1	0,45	0.000110	18
Candidate2	0,49	0.003474	17
Candidate3	0,50	0.002120	16
Candidate4	0,45	0.000110	18
Candidate5	0,45	0.000110	18
Candidate6	0,45	0.000110	18
Candidate7	0,62	0.002395	4
Candidate8	0,61	0.003479	9
Candidate9	0,53	0.002871	15
Candidate10	0,45	0.000110	18
Candidate11	0,45	0.000110	18
Candidate12	0,62	0.005833	1

Candidate13	0,62	0.002074	6
Candidate14	0,61	0.002142	11
Candidate15	0,53	0.002298	14
Candidate16	0,45	0.000110	18
Candidate17	0,62	0.005833	1
Candidate18	0,59	0.004511	12
Candidate19	0,61	0.002944	7
Candidate20	0,61	0.002715	10
Candidate21	0,45	0.000110	18
Candidate22	0,62	0.005833	1
Candidate23	0,59	0.004993	13
Candidate24	0,62	0.002149	5
Candidate25	0,61	0.003157	8

Visualisasi rata-rata skor tes dalam tabel 6 dapat dilihat melalui grafik yang ditunjukkan pada gambar 5 berikut.



Gambar 5. Grafik rata-rata skor tes

Selain itu, validasi silang K-fold yang dijalankan pada program penelitian ini adalag n_splits=5 dengan n_repeats=3. Sehingga masing-masing tes split tersebut menghasilkan nilai seperti tabel 7 berikut.

Tabel 7. *Split Test Score*

Candidate	K-fold split data				
	Ke-1	Ke-2	Ke-3	Ke-4	Ke-5
1	0,45	0,45	0,45	0,45	0,45
2	0,48	0,49	0,48	0,49	0,49
3	0,49	0,50	0,49	0,50	0,50
4	0,45	0,45	0,45	0,45	0,45
5	0,45	0,45	0,45	0,45	0,45
6	0,45	0,45	0,45	0,45	0,45
7	0,63	0,62	0,62	0,62	0,62
8	0,62	0,61	0,61	0,61	0,61
9	0,53	0,53	0,53	0,54	0,53
10	0,45	0,45	0,45	0,45	0,45
11	0,45	0,45	0,45	0,45	0,45
12	0,63	0,62	0,61	0,63	0,62
13	0,62	0,62	0,62	0,62	0,62
14	0,61	0,61	0,61	0,61	0,61
15	0,54	0,53	0,54	0,54	0,53
16	0,45	0,45	0,45	0,45	0,45
17	0,63	0,62	0,61	0,63	0,62
18	0,59	0,58	0,59	0,60	0,59
19	0,62	0,61	0,61	0,61	0,62
20	0,61	0,61	0,61	0,61	0,61
21	0,45	0,45	0,45	0,45	0,45
22	0,63	0,62	0,61	0,63	0,62
23	0,59	0,58	0,58	0,59	0,59
24	0,62	0,62	0,62	0,62	0,62
25	0,62	0,61	0,61	0,61	0,61

Berdasarkan informasi pada data tabel 7 tersebut, terlihat bahwa kandidat 22, kandidat 17 dan kandidat 12 memperoleh nilai yang lebih baik pada *split test score* ke-4 yakni 0,629211.

Kesimpulan

Berdasarkan hasil pengujian terhadap prediksi struktur sekunder protein menggunakan metode SVM, dapat disimpulkan bahwa prediksi struktur sekunder protein dilakukan dengan cara mengambil urutan ganjil untuk struktur primer dan urutan genap untuk struktur sekunder. Dataset yang digunakan dalam penelitian ini merupakan data RS126 yang memiliki 150 instance dan setiap instance berisi 2 baris.Baris pertama = struktur primer (asam amino) dan Baris kedua = struktur sekunder (C, H, E).

Optimalisasi parameter dan hasil prediksi struktur sekunder protein pada algoritma SVM ini diperoleh dengan proses validasi silang pada data latih dan data uji dengan melakukan perhitungan kombinasi antara masing-masing nilai parameter, estimator, dan hasil split k-fold data.

Hasil optimalisasi menunjukkan bahwa parameter pada kandidat 22,17, dan 12 juga merupakan parameter terbaik selain parameter C=1 dan gamma = 0,1 pada penelitian sebelumnya dalam memprediksi struktur sekunder protein.

Pada penelitian selanjutnya dapat dilakukan optimasi algoritma (*hyperparameter tuning*) pada algoritma *Support Vector Machine* dengan menggunakan metode *grid search optimization* dengan menambahkan dua jenis

kernel lainnya selain kernel *radial basis function*, yakni kernel linear dan polynomial.

Referensi

Amalia, D. H., & Yustanti, W. (2021). Klasifikasi Buku Menggunakan Metode Support Vector Machine pada Digital Library. *Journal of Informatics and Computer Science (JINACS)*, *3*(01), 55–61. https://doi.org/10.26740/jinacs.v3n01.p55-61

Haryanto, T., & Budiman, B. (2015). Penggunaan Fitur Kimia Fisik dan Posisi Atom untuk Prediksi Struktur Sekunder Protein. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *1*(2). https://doi.org/10.26418/jp.v1i2.11919

Hermanto, H., Mustopa, A., & Kuntoro, A. Y. (2020). Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *5*(2), 211–220. https://doi.org/10.33480/jitk.v5i2.1181

Isnain, A. R., Sakti, A. I., Alita, D., & Marga, N. S. (2021). Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma SVM. *JDMSI*, 31-37.

Kurniawati, P., & Banowati, R. (2018). Modul Asam Amino, Peptida dan Protein. *Diploma Chemistry Uii*, 15–35.

Nanda, S., Arjunan, V. E. L., & Deris, S. (2007). *Prediction of protein secondary structure*. *35*(C), 81–90.

Prima, J., Sistem, J., Komputer, I., No, V., Radhi, M., Ryan, D., Sitompul, H., Sinurat, S. H., & Indra, E. (2021). *PREDIKSI HARGA MOBIL MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING*. *4*(2), 1–5.

Purnami, W., Regresi, A. M., & Ordinal, L. (2012). Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM) . *Jurnal Sains dan Seni ITS*, D130-D135.

Rarasmaya Indraswari, A. Z. (2017). RBF Kernel Optimization Method With Particle SWARM Optimization On SVM Using The Analysis Of Input Data's Movement. *Jurnal Ilmu Komputer dan Informasi*, 36-42.

Sari, D. P., & Haryanto, T. (2016). *Penerapan Algoritme Viterbi pada Hidden Markov Model (HMM) untuk Prediksi Struktur Sekunder Protein Penerapan Algoritme Viterbi pada Hidden Markov Model (HMM) untuk Prediksi Struktur Sekunder Protein*. *February*.

Sistem, R., Kartini, D., Abadi, F., & Saragih, T. H. (2021). *Prediksi Tinggi Permukaan Air Waduk Menggunakan Artificial Neural*. *1*(10), 39–44.

S, S., & K, M. (2019). A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 219-230.

Sutanto, V. M., Sukma, Z. I., & Afiahayati, A. (2020). Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine. *International Journal of Intelligent Engineering and Systems*, *14*(1), 232–243.

Tasari, A., Tarigan, D. D., Nia, E., Br, D., & S, K. S. (2022). *Perbandingan Algoritma Support Vector Machine dan KNN dalam Memprediksi Struktur Sekunder Protein*. *9*(2), 172–179.

Wenwen, L., Xing, X., Liu, F., & Yu Zhang. (2014). Application of improved grid search algorithm on SVM for classification of tumor gene. *International Journal of Multimedia and Ubiquitous Engineering*, *9*(11), 181–188. https://doi.org/10.14257/ijmue.2014.9.11.18.

Wulandari, A. (2020). Aplikasi Support Vector Machine (SVM) untuk Pencarian Binding Site Protein-Ligan. *MATHunesa: Jurnal Ilmiah Matematika*, *8*(2), 157–161. https://doi.org/10.26740/mathunesa.v8n2.p157-161.

Yan, T., Shen, S., Zhou, A., & Chen, X. (2022). Journal of Rock Mechanics and Geotechnical Engineering Prediction of geological characteristics from shield operational parameters by integrating grid search and K -fold cross validation into stacking classi fi cation algorithm. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(4), 1292–1303.

Yasin, H., Caraka, R. E., & Hoyyi, A. (2016). *Prediction of Crude Oil Prices using Support Vector Regression (SVR) with*

grid search – cross validation algorithm.
12(4), 3009–3020.