Imputasi *k*NN pada *Hybrid Support Vector Machine* untuk Kasus Klasifikasi Tingkat Keparahan COVID-19

Sayyid Nur Cahyo Abdul Jalil[1], Shofi Andari[2], dan Santi Wulan Purnami[3]
Departemen Statistika, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111, Indonesia
e-mail: cahyo.19062@mhs.its.ac.id[1], shofi.andari@statistika.its.ac.id[2], santi_wp@statistika.its.ac.id[3]

***Abstrak*— COVID-19 merupakan penyakit infeksi pada organ-organ pernapasan yang disebabkan oleh virus SARS-CoV-2 dan telah menyebar ke seluruh dunia sejak 2019. Dalam mengurangi akibat yang mungkin disebabkan oleh COVID-19, diperlukan diagnosis yang lebih akurat dalam membedakan pasien COVID-19 berdasarkan tingkat keparahannya sehingga dapat dijadikan sebagai *early warning* dalam memberikan perawatan yang tepat pada pasien. Pada permasalahan klasifikasi pada bidang medis, terdapat kemungkinan adanya *missing value* dikarenakan rekam medis yang tidak tercatat sepenuhnya akibat kesalahan pencatatan data rekam medis yang dapat menyebabkan *overfitting*. Penelitian ini bertujuan untuk mengembangkan model prediksi yang akurat dan dapat menangani *overfitting* menggunakan *Hybrid Random Forest-Support Vector Machine* (RF-SVM) serta mengatasi *missing value* yang terdapat pada data rekam medis pasien COVID-19 menggunakan imputasi *k*NN. Hasil analisis pada penelitian ini menunjukkan bahwa sebagian besar pasien yang terjangkit COVID-19 menunjukkan risiko yang rendah serta memiliki karakteristik pada tingkat keparahan rendah antara lain pasien berusia muda, tidak pernah menderita diabetes dan hipertensi, laju pernafasan normal/rendah, saturasi oksigen tinggi, dan tidak menunjukkan gejala sesak nafas. Penggunaan imputasi *k*NN menunjukkan bahwa metode ini mampu menghasilkan performa prediksi yang lebih besar dibandingkan metode penanganan *missing value* mean/modus. Penggunaan metode *hybrid* RF-SVM menunjukkan bahwa metode ini mampu secara efektif mengurangi *overfitting* dan meningkatkan kemampuan generalisasi model. Hasil permodelan menggunakan imputasi *k*NN dan *hybrid Random Forest – Support Vector Machine* menunjukkan performa F1-*score* mencapai 97,6% dan nilai AUC sebesar 98% dalam memprediksi data baru.**

***Kata Kunci*—COVID-19, *Hybrid Support Vector Machine*, Imputasi *K*NN, Klasifikasi.**

I. PENDAHULUAN

COVID-19 merupakan penyakit infeksi pada organ-organ pernapasan yang disebabkan oleh virus SARS-CoV-2 [22]. Penyakit ini pertama kali muncul di Wuhan, Tiongkok pada tahun 2019 dan sejak itu telah menyebar ke seluruh dunia, menyebabkan pandemi global. Sejak awal penyakit ini terdeteksi pada Desember 2019 sampai 29 Oktober 2021, tercatat sebanyak 249 juta orang di seluruh dunia terpapar COVID-19 [22]. Pada skala nasional, Indonesia mengalami lonjakan kasus COVID-19 yang sangat tajam pada tahun 2021 di mana terdapat total lebih dari 4 juta kasus per 31 Oktober 2021 dengan tingkat kematian tertinggi di Indonesia mencapai 3,38% [17]. Dalam mengurangi akibat yang mungkin disebabkan oleh COVID-19, diperlukan diagnosis yang lebih akurat dalam membedakan pasien yang memerlukan perawatan intensif, pasien yang diperbolehkan melakukan isolasi mandiri hingga pasien yang berkemungkinan meninggal. Terdapat beberapa permasalahan yang umum ditemui dalam kasus klasifikasi pasien, antara lain terdapat kemungkinan kesalahan pencatatan oleh petugas medis sehingga menyebabkan adanya nilai data yang tidak tercatat oleh rumah sakit serta ketidakpastian pengukuran dan pencatatan data yang memungkinkan variabilitas serta *noise* pada data. Penelitian ini berfokus pada penanganan data yang hilang (*missing value*) serta penanganan *overfitting* yang disebabkan variabilitas data rekam medis pada permasalahan tingkat keparahan pasien COVID-19 dengan studi kasus di Rumah Sakit Universitas Airlangga Surabaya.

Penelitian terkait klasifikasi COVID-19 telah banyak dilakukan sebelumnya untuk mendeteksi COVID-19 berdasarkan hasil dari uji darah serta gejala yang ditunjukkan oleh pasien COVID-19 [2&24]. Penelitian tersebut menggunakan berbagai metode prediktif *machine learning*, salah satunya yaitu *k-Nearest Neighbor* (*k*NN) dan *Support Vector Machine* (SVM) dalam memprediksi tingkat keparahan pasien. Penelitian tersebut menghasilkan bahwa model *machine learning* SVM menunjukkan performa yang baik dalam memprediksi keparahan penyakit COVID-19 dari pasien berdasarkan data rekam medis. Penelitian lain dalam mengelompokkan keparahan pasien COVID-19 juga dilakukan menggunakan *Combine Sampling Support Vector Machine* (SVM) dengan studi kasus di Kota Surabaya [13]. Beberapa penelitian sebelumnya yang pernah dilakukan mengenai penerapan algoritma imputasi *k*NN dalam menangani permasalahan *missing value* yang mengimplementasikan imputasi *k*NN dibandingkan dengan metode imputasi lain dalam permasalahan klasifikasi kelulusan, prediksi cuaca dan monitoring kualitas air [9,14&20]. Beberapa penelitian tersebut menunjukkan bahwa penggunaan imputasi *k*NN mampu menunjang performa model prediktif serta performa dan stabilitas imputasi *k*NN yang baik dalam mereplikasi data dengan persentase data yang hilang berbeda-beda.

Permasalahan *overfitting* dapat diantisipasi dengan penerapan metode *ensemble* yang menggunakan konsep *random sampling with replacement* pada proses pembentukan model sehingga pada penelitian diusulkan metode *Hybrid Random Forest-Support Vector Machine* (RF-SVM) dalam mengantisipasi *overfitting* dan menghasilkan model yang dapat digeneralisasi terbilang cukup efektif. Penerapan algoritma prediktif menggunakan model *hybrid* ini telah dilakukan dengan menerapkan algoritma *hybrid Random Forest-Support*

Vector Machine (RF-SVM) untuk peningkatan performa dan generalisir model pada kasus klasifikasi identifikasi suara, dan identifikasi objek [11&15]. Beberapa penelitian tersebut menunjukkan bahwa penerapan metode RFSVM dengan semakin banyak subset data yang terbentuk akan memiliki prediksi yang *robust* terhadap data baru serta memiliki performa klasifikasi lebih baik daripada model individual SVM.

Penyakit COVID-19 dapat bervariasi dalam tingkat keparahannya, mulai dari gejala ringan hingga gejala yang sangat berat yang dapat menyebabkan kematian. Penelitian ini bertujuan untuk mendeskripsikan karakteristik data rekam medis pasien pada data hasil imputasi *k*NN dan memperoleh performa akurasi klasifikasi yang optimal untuk kasus tingkat keparahan pasien COVID-19 menggunakan imputasi *k*NN dan *hybrid* RF-SVM pada data hasil imputasi. Perlakuan yang digunakan dalam penelitian ini adalah imputasi *k*NN dengan perhitungan parsial pada data kategorik dan numerik untuk mereplikasi data yang hilang serta penggunaan metode *hybrid* RF-SVM dalam mengurangi efek *overfitting* pada hasil prediksi. Diharapkan penelitian ini mampu meningkatkan performa dan mengatasi *overfitting* dengan model prediktif *hybrid* serta mengatasi permasalahan *missing value* pada data rekam medis dengan metode imputasi *k*NN pada studi kasus pasien COVID-19 di RS UNAIR sehingga dapat membantu pihak medis untuk memberikan *treatment* yang lebih sesuai pada keseluruhan pasien COVID-19.

II. METODE PENELITIAN

A. Imputasi kNN

Permasalahan nilai data yang hilang secara statistik dapat dibagi menjadi tiga kategori berdasarkan keacakannya, yaitu hilang sepenuhnya acak (*missing completely at random*), hilang acak (*missing at random*), dan hilang tidak acak (*missing not at random*) [8]. *Missing completely at random* (MCAR) adalah kondisi data yang hilang murni secara acak tanpa ada korelasi dengan variabel lainnya, dimana data yang hilang dapat dihapus pada analisis karena tidak memuat informasi khusus di dalamnya. *Missing at random* (MAR) ialah kondisi data yang hilang secara acak memiliki korelasi dengan variabel lain namun tidak berkorelasi dengan variabel dengan nilai data yang hilang itu sendiri, di mana ditangani dengan metode imputasi seperti mean, modus, atau imputasi berbasis model prediksi. *Missing not at random* (MNAR) adalah kondisi data yang hilang tidak secara acak dan berkaitan dengan variabel dengan data yang hilang itu sendiri yang ditangani dengan metode imputasi dengan mempertimbangkan permodelan nilai data seperti imputasi *k*NN. Penentuan keacakan *missing value* ditentukan dengan uji *little's* MCAR dengan rumus berikut.

$$d_0^2 = \sum_{s}^{S} n_s \left(\bar{\mathbf{x}}_{\mathbf{o}_s} - \boldsymbol{\mu}_{\mathbf{o}_s}\right)^T \sum\nolimits_{\mathbf{o}_s}^{-1} \left(\bar{\mathbf{x}}_{\mathbf{o}_s} - \boldsymbol{\mu}_{\mathbf{o}_s}\right)$$
$$df = \sum_{s=1}^{S} p_s - p \quad (1)$$

Keterangan : S = pola *missing value*

$\bar{\boldsymbol{x}}_{\boldsymbol{O}_s}$ = sampel data **x** dengan indeks $\mathbf{o_s}$

$\boldsymbol{\mu}_{\boldsymbol{O}_s}$ = rata-rata dari data **x** yang terobservasi

$\sum_{o_s}^{-1}$ = invers matriks kovarian pada pola s

Tahapan pengolahan algoritma imputasi *k*NN secara umum sebagai berikut.

1. Menentukan nilai *k* jumlah pengamatan terdekat.
2. Menghitung jarak antara nilai yang hilang dengan data *training* secara parsial menggunakan *Euclidean* dan *Hamming Distance* [10] dengan rumus berikut.

$$Dist_E\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{\sum_{m=1}^{M}\left(x_{im} - x_{jm}\right)^2} \quad (2)$$

$$Dist_H\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{m=1}^{M} 1\left(x_{im} \neq x_{jm}\right) \quad (3)$$

Keterangan :

x_{im} = skalar data dengan *missing value*

x_{jm} = skalar data pengamatan ke-j

m = indeks variabel/fitur

$1\left(x_{im} \neq x_{jm}\right)$ = fungsi indikator dengan nilai 1 jika $x_{im} \neq x_{jm}$ dan bernilai 0 jika $x_{im} = x_{jm}$.

3. *Distance* yang diperoleh dilakukan fungsi *agregat* untuk menggabungkan jarak *Euclidean* dan *Hamming*.
4. Menghitung nilai rata-rata jarak dari *k* nilai data terdekat untuk nilai variabel diimputasi x_i numerik dan data kategorikal yang hilang direplikasi menggunakan nilai mayoritas dari *k* nilai data terdekat yaitu x_{knn} [25].

$$x_{im} = \begin{cases} \arg\max_v \left(\sum_{D_k} 1\left(x_{knn,im} = v\right) \right), & \text{Kategori} \\ \frac{1}{k}\sum_{i=1}^{k} x_{knn,im}, & \text{Numerik} \end{cases} \quad (4)$$

Keterangan :

D_k = $\left\{\left(\mathbf{x}_j, \mathbf{x}_i\right)\right\}$ = himpunan *k nearest neighbor*

I = {1,2,…,k}

V = nilai dalam domain fitur variabel yang diimputasi x_i

$1(x_{knn,im}=v)$ = fungsi indikator dengan nilai 1 jika argumennya benar dan 0 ketika argumennya salah.

5. Mengisi nilai *missing value* menggunakan nilai *output* pada langkah 4.
6. Hasil data yang telah diimputasi normalisasi *Z-score* pada data numerik dengan persamaan sebagai berikut.

$$Z - score = \frac{(x - \mu)}{\sigma} \quad (5)$$

Keterangan =

μ = mean dari kumpulan data x

σ = standar deviasi dari kumpulan data x

Metode imputasi *k*NN merupakan metode yang fleksibel untuk data diskrit maupun kontinu serta dapat digunakan untuk penanganan *missing value* secara akurat [7].

B. Combine Sampling

Penggunaan *Combine Sampling* atau metode *hybrid resampling*, baik data *oversampling* (SMOTE) dan data undersampling (Tomek Links) diintegrasikan bersama untuk menghasilkan dataset tunggal.

1) Syntetic Minority Over-sampling Technique

SMOTE merupakan algoritma dalam mengatasi *imbalanced datasets* dengan metode *oversampling*. SMOTE akan menghasilkan data sintetik yang dibuat dari data dalam sampel kelas minoritas. Metode ini pertama kali diperkenalkan oleh Chawla [5]. Tahapan perhitungan algoritma SMOTE sebagai berikut.

1. Setiap data pada kelas minoritas akan direplikasi untuk mencari tetangga terdekat dengan menggunakan pengukuran jarak antara titik data, di mana pada penelitian ini akan digunakan jarak *Euclidean* dan jarak *Hamming* pada Persamaan 2 dan 3.
2. Menghitung data sintetis menggunakan persamaan berikut.

$$\mathbf{x_{syn}} = \mathbf{x_i} + \left(\mathbf{x_{knn,i}} - \mathbf{x_i}\right)\delta \qquad (6)$$

dengan merupakan nilai acak dari 0 hingga 1 [13].

2) Tomek Links

Metode Tomek Links adalah salah satu teknik *undersampling* yang digunakan untuk mengatasi masalah kelas tidak seimbang dalam data. Metode ini bekerja dengan menghapus titik data dari kelas mayoritas yang memiliki jarak dekat dengan titik data dari kelas minoritas yang sama. Hal ini dilakukan dengan tujuan untuk meningkatkan keakuratan dan konsistensi klasifikasi data pada kelas minoritas [21].

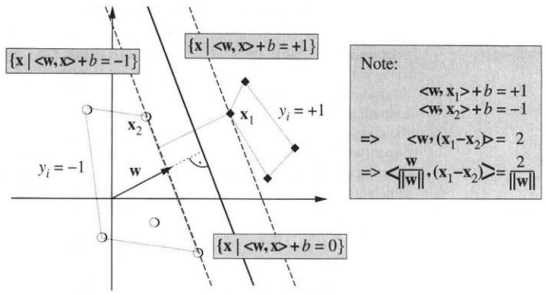
C. Hybrid Support Vector Machine

Model yang digunakan menggabungkan konsep *ensemble* yaitu *Random Forest* yang dapat mengurangi variabilitas pada hasil prediksi sehingga generalisasi yang diperoleh menjadi jauh lebih baik pada data *out-sample*.

1) Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah sebuah algoritma *supervised learning* yang dirumuskan oleh Vladimir Vapnik pada tahun 1992. Konsep dasar dari SVM adalah mencari sebuah *hyperplane* (bidang n-dimensi) yang mampu memisahkan dua kelas data dengan margin terbesar yang memungkinkan [19]. Permasalahan paling mendasar pada logika klasifikasi berupa label respon biner, dan jika diasumsikan data *training* sebagai himpunan $\mathbf{D} = [(x_1,y_1),\ldots,(x_n,y_n)]$ di mana respon Y={-1,1} dengan vektor pembobot $\mathbf{w}^T = [w_1,w_2,\ldots,w_d] \in \mathbb{R}^d$, variabel prediktor $\mathbf{x}^T = [x_1,x_2,\ldots,x_d] \in \mathbb{R}^d$ dan angka riil $b \in \mathbb{R}$ sebagai skalar, maka kriteria data pada klasifikasi biner dengan <..> merupakan *dot product* dituliskan berikut [18].

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i \rangle + b > 0, \quad & y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b < 0, \quad & y_i = -1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b = 0, \quad & \textit{hyperplane} \end{aligned} \qquad (7)$$



Gambar 1. *Hyperplane* Klasifikasi Biner SVM

Dalam kasus di mana data tidak dapat dipisahkan secara linier, diperlukan fungsi khusus *kernel* dalam memetakan data ke dalam ruang fitur yang memiliki dimensi yang lebih tinggi, sehingga dapat memisahkan data yang tidak dapat dipisahkan secara linier dalam dimensi asli mereka [13]. Terdapat beberapa fungsi *kernel* dalam algoritma SVM, salah satunya adalah *kernel Radial Base Function* (RBF), dikenal juga sebagai fungsi *Gaussian RBF*, fungsi ini mengukur jarak antara dua data menggunakan fungsi *Gaussian*, dengan perumusan sebagai berikut.

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) \qquad (8)$$

Dengan $\gamma > 0$ merupakan parameter gamma yang menunjukkan seberapa jauh pengaruh dari sampel pada *training data*. Semakin kecil parameter *gamma*, maka nilai data yang dipertimbangkan untuk membentuk *hyperplane* semakin banyak.

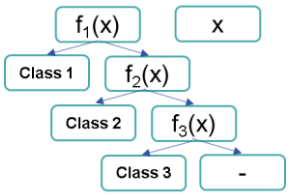
Penggunaan metode SVM pada awalnya didasarkan pada prediksi data biner, sehingga lebih lanjut dikembangkan metode pendekatan untuk memprediksi permasalahan bertipe *multiclass*. Terdapat 2 pendekatan (*decision function*) SVM-*multiclass* pada umumnya, antara lain [3].

1. *One-Against-All* (OAA) / *One-Versus-Rest* (OVR)

Pendekatan ini memprediksi *output* dengan membuat fungsi prediksi sebanyak jumlah *class* dalam variabel respons. Fungsi tersebut membandingkan antara respons *class* ke-*a* dengan gabungan *class* non-*a*. Pendekatan ini menghitung nilai fungsi objektif dalam menentukan *hyperplane* dituliskan sebagai berikut [3].

$$\min\left\{\frac{1}{2}\left(\mathbf{w}^a\right)^T \mathbf{w}^a + c\sum_{j=1}^{s}\xi_j^a\right\} \qquad (9)$$

Di mana *a* merupakan *class* respons, *s* adalah banyak data pada *training* model, ξ adalah variabel *slack* sebagai error dalam prediksi, serta *c* ialah parameter *trade-off* antara margin dan kesalahan klasifikasi, disebut juga parameter regularisasi. Ilustrasi pengambilan keputusan algoritma ini ditampilkan pada Gambar 2.



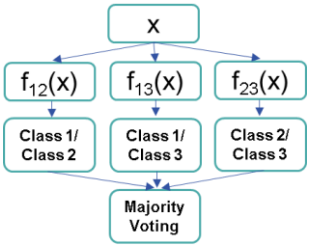
Gambar 2. Diagram Alur *One-Against-All*

2. *One-Against-One* (OAO) / *One-Versus-One* (OVO)

Pendekatan ini memprediksi *class* respons dengan membandingkan data dalam *class* ke-*a* dengan data pada kelas lain *b*. Dalam algoritma ini, jika terdapat sebanyak *h* jumlah *class* dalam respons maka fungsi yang akan terbentuk sebanyak *h*(*h-1*)/2. Fungsi objektif pendekatan ini dirumuskan sebagai berikut [3].

$$\min\left\{\frac{1}{2}\left(\mathbf{w}^{ab}\right)^{T}\mathbf{w}^{ab}+c\sum_{j=1}^{s}\xi_{j}^{ab}\right\} \qquad (10)$$

Dengan *a* merupakan *class* respons, diagram pengambilan keputusan pada algoritma ini ditampilkan pada Gambar 3.



Gambar 3. Diagram Alur *One-Against-One*

2) Random Forest (RF)

Random Forest merupakan kombinasi sebanyak *t* model prediktif dari *Decision Tree* (DT) di mana setiap model *Decision Tree* bergantung pada nilai subset data (*bootstrap*) acak yang diambil sampelnya secara independen dengan distribusi yang sama [1]. Jika terdapat data *training* sebagai himpunan $\mathbf{D} = [(x_1,y_1),\ldots,(x_n,y_n)]$ dengan banyak data adalah *n*, *bootstrap* akan membentuk *t* subset data *training* baru $\mathbf{D_i}$ dari $\mathbf{D_1}, \ldots, \mathbf{D_i}, \ldots, \mathbf{D_t}$ sebanyak *n* pengamatan dengan setiap $\mathbf{D_i}$ adalah sampel acak *uniform* dengan penggantian. Langkah kerja algoritma *Random Forest* sebagai berikut [16].

1. Dengan *t* jumlah pembagian subset data, maka *i* dari $1,\ldots,i,\ldots,t$:
   - Tentukan sampel $\mathbf{D_i}$ dengan *random sampling with replacement* dari **D** data *training*.
   - Buat model *Decision Tree* menggunakan subset data $\mathbf{D_i}$ dengan perulangan setiap *terminal node*.
     i) Pilih variabel acak dari seluruh variabel.
     ii) Pilih variabel terbaik sebagai *split point* menggunakan nilai *gini impurity*.

$$Gini(n)=1-\sum_{a=1}^{A}P_a^2 \qquad P_a=\frac{n_a}{n} \qquad (11)$$

dengan *A* : kelas label respon
P_a : probabilitas kelas ke-*a*
n_a : jumlah pengamatan ke-*a*

iii) Bagi *node* yang menghasilkan 2 *node* baru. Setelah dilakukan *split*, perhitungan *gini* dirumuskan:

$$Gini_M(n)=\frac{|n_1|}{|n|}Gini(n_1)+\frac{|n_2|}{|n|}Gini(n_2) \qquad (12)$$

dengan *M* : Variabel dengan *split* biner.

2. *Output* dari kumpulan model *Decision Tree* $\{\mathbf{D_1},\ldots,\mathbf{D_i},\ldots,\mathbf{D_t}\}$ dilakukan *majority voting* yaitu cara penentuan hasil prediksi dengan menggunakan pilihan mayoritas terhadap hasil yang muncul. Voting mayoritas pada kasus klasifikasi menggunakan modus dari tiap *output* subset data yang berupa label kategorik.

3) Hybrid RF-SVM Multiclass

RF-SVM (*Random Forest-Support Vector Machine*) adalah suatu metode penggabungan antara *Random Forest* dan *Support Vector Machine* untuk memecahkan masalah klasifikasi yang rumit. Metode ini bertujuan untuk memperoleh model klasifikasi yang memiliki kemampuan generalisir yang baik dan performa model yang akurat dalam memprediksi data baru. Langkah kerja algoritma RF-SVM sebagai berikut [15].

1. Menentukan data *training* dan *testing*, gunakan data *training* sebagai *input* dalam pembagian subset data.
2. Menentukan *t* jumlah pembagian subset dan bagi data *training* sebanyak *t* subset data. Pembagian subset data dilakukan dengan konsep *random sampling with replacement*.
3. Untuk setiap subset data *training* $\mathbf{D_i}$, buat model prediksi SVM *multiclass*.
4. Menentukan hasil prediksi dari setiap model SVM pada subset data *training* terhadap data *testing*.
5. Menentukan hasil prediksi RF-SVM dengan *majority voting* pada hasil prediksi tiap subset data.
6. Menghitung kebaikan model berdasarkan data *testing* dan hasil prediksi.

D. Grid Search Algorithm

Algoritma *Support Vector Machine* (SVM) merupakan model prediksi *machine learning* yang sensitif terhadap perubahan parameternya dalam menentukan batas pemisah (*hyperplane*) pada permasalahan klasifikasi/regresi. *Grid search* merupakan metode pengoptimalan parameter yang baik dalam menemukan solusi optimal dengan membentuk model terbaik berdasarkan himpunan parameter yang diberikan [12].

E. Q-fold Cross Validation

Q-fold cross validation merupakan teknik untuk membagi data *training* dan data *testing* menjadi sejumlah *q* grup, dengan $q \geq 2$ dan menggunakan setiap grup sebagai *testing datasets* untuk mengevaluasi model [26]. Hal ini dilakukan untuk melihat apakah model masih *reliable* untuk data uji yang berbeda-beda.

F. Evaluasi Model

Evaluasi model dilakukan dengan memprediksi kebaikan model pada setiap *fold* yang terbentuk. Model prediktif dalam permasalahan klasifikasi *multi-class* secara umum dapat dievaluasi menggunakan *confusion matrix* yang merepresentasikan hasil prediksi dengan nilai yang sebenarnya [4]. Pada permasalahan klasifikasi *multi-class* akan dihasilkan *confusion matrix* pada Tabel 1 [23].

Tabel 1.
Confusion Matrix

		Nilai Prediksi				*Row Margin*
		1	2	⋯	A*	
Nilai Aktual	1	n_{11}	n_{12}	⋯	n_{1A}^*	n_1
	2	n_{21}	n_{22}	⋯	n_{2A}^*	n_2
	⋮	⋮	⋮	⋱	⋮	⋮
	A	n_{A1}	n_{A2}	⋯	n_{AA}^*	n_A
Column Margin		n_1	n_2	⋯	n_A^*	N

dengan A = {1,2,…,A} merupakan himpunan label *multiclass* sebagai nilai aktual dan A* = {1,2,…,A*} merupakan himpunan label *multiclass* sebagai nilai hasil prediksi.
Misal jika dilakukan perhitungan pada label kelas 1 akan memiliki kriteria berikut.
- *True Positive* = n_{11}
- *True Negative* = $(n_2 - n_{21}) + (n_3 - n_{31}) + \ldots + (n_A - n_{A1})$
- *False Negative* = $n_{12} + n_{13} + \ldots + n_{1A}^*$
- *False Positive* = $n_{21} + n_{31} + \ldots + n_{A1}$

Menggunakan kriteria diatas dapat dilakukan beberapa perhitungan yang menunjukkan kualitas model sebagai berikut.
1. Akurasi merupakan proporsi dari prediksi yang benar.
$$Akurasi = \frac{TP + TN}{FP + FN + TP + TN} \quad (13)$$
2. *F-score* merupakan perhitungan yang menggabungkan presisi dan sensitivitas dalam mengevaluasi kebaikan model.
$$F - score = \frac{presisi \times sensitivitas}{presisi + sensitivitas} \quad (14)$$
3. *Area Under Curve* (AUC) merupakan statistik yang menghitung area dari kurva *Receiver Operating Charateristic* (ROC). Kurva ROC merepresentasikan sensitivitas dengan 1-spesifisitas. Kurva ROC umumnya digunakan untuk menggambarkan keakuratan prediksi dan menentukan nilai *cut-off* yang optimal.
$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (15)$$

G. Uji Independensi

Uji independensi dilakukan untuk melihat apakah terdapat pengaruh dari suatu variabel terhadap variabel lain melalui persebaran datanya.

1) Uji Chi-Square

Uji *chi-square* merupakan salah satu pengujian secara statistik dalam melihat pengaruh dari satu variabel kategori terhadap variabel kategori lainnya berdasarkan nilai dari tabel kontingensi antar variabel. Hipotesis awal uji *chi-square* adalah dua kriteria merupakan independen. Statistik uji *chi-square* dihitung berdasarkan persamaan berikut [6].
$$X^2 = \sum_{b=1}^{B}\sum_{c=1}^{C}\frac{\left(O_{bc} - E_{bc}\right)^2}{E_{bc}} \quad (16)$$
$$df = (B-1)(C-1)$$
Keterangan : B = banyak baris pada tabel kontingensi
C = banyak kolom pada tabel kontingensi
O_{bc} = nilai tabel kontingensi pada indeks b,c
E_{bc} = ekspektasi pada indeks b, c = $\frac{n_b \times n_j}{n}$
n = banyak pengamatan

Keputusan dari uji statistik ini akan menolak hipotesis awal atau dua kriteria sampel bersifat dependen apabila nilai *chi-square* hitung > χ^2_{df} atau *p-value* < α .

2) Uji Kruskal-Wallis

Uji *kruskal-wallis* digunakan untuk melihat pengaruh dari variabel kategori terhadap variabel numerik dengan melihat dari perubahan mediannya di tiap label faktor berbeda. Hipotesis awal yang digunakan adalah fungsi distribusi populasi identik dimana median sama untuk semua faktor, yang dapat diartikan faktor tidak berpengaruh terhadap nilai dari variabel yang diamati. Statistik uji *kruskal-wallis* dihitung berdasarkan persamaan berikut [6].
$$H = \frac{12}{n(n+1)}\sum_{l=1}^{L}\frac{R_l}{n_l} - 3(n+1) \quad (17)$$
$$df = L - 1$$
Keterangan : n = banyak pengamatan
L = label pada faktor
R = jumlah dari *rank* untuk pengamatan ke *l*

Keputusan dari uji statistik ini akan menolak hipotesis awal atau dua kriteria sampel memiliki median tidak sama setidaknya satu apabila didapati nilai H > χ^2_{df} atau *p-value* < α .

H.Tingkat Keparahan COVID-19

Pada pasien COVID-19 terindikasi tingkat keparahan penyakit ringan (40%) atau sedang (40%), serta sekitar 15% berkembang menjadi penyakit parah, dan 5% memiliki penyakit kritis [22]. Tingkat keparahan tersebut memiliki gejala yang berbeda-beda, contohnya untuk pasien dengan gejala ringan mengalami demam, batuk, nyeri tenggorokan, dan hidung tersumbat. Pada pasien dengan gejala sedang memiliki gejala yang tidak jauh berbeda dengan gejala ringan namun pasien memiliki gejala yang lebih akut dan mengalami sesak nafas. Pasien dengan tingkat keparahan berat menunjukkan demam yang tinggi >38^0C, ISPA, pneumonia, hingga saturasi oksigen yang sangat rendah. Pada tingkat kritis, pasien memiliki kemungkinan meninggal tinggi menunjukkan gejala yaitu gagal organ, gagal pernafasan, hingga syok [13].

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan merupakan data sekunder rekam medis pasien COVID-19 pada periode gelombang kedua pandemi di Indonesia pada waktu Mei hingga Oktober 2021. Data penelitian merupakan data sekunder rekam medis pasien dengan studi kasus pasien COVID-19 di Rumah Sakit Universitas Airlangga dengan 668 pasien dan 16 variabel pada Tabel 2.

B. Variabel Penelitian

Variabel yang digunakan dijelaskan pada Tabel 2.

Tabel 2.
Variabel Penelitian

Variabel	Nama Variabel	Keterangan	Skala
Y	Tingkat Keparahan	0 : Meninggal 1 : Risiko berat 2 : Risiko ringan hingga sedang	Ordinal
X_1	Usia	Usia pasien ketika dirawat	Rasio
X_2	Jenis Kelamin	0 : Perempuan 1 : Laki-laki	Nominal
X_3	Diabetes Melitus	0 : Tidak memiliki riwayat diabetes 1 : Memiliki riwayat diabetes	Nominal
X_4	Hipertensi	0 : Tidak punya riwayat hipertensi 1 : Memiliki riwayat hipertensi	Nominal
X_5	Gagal Ginjal Kronis	0 : Tidak memiliki riwayat gagal ginjal kronis 1 : Memiliki riwayat gagal ginjal	Nominal
X_6	Penyakit Jantung	0 : Tidak memiliki riwayat penyakit jantung 1 : Memiliki riwayat penyakit jantung	Nominal
X_7	*Respiratory Rate*	*Respiratory rate* (nafas per menit)	Rasio
X_8	SpO_2	Saturasi oksigen/konsentrasi oksigen dalam darah (%)	Rasio
X_9	Tekanan Darah Sistolik	Tekanan darah tertinggi pasien dalam perawatan saat jantung konstraksi	Rasio
X_{10}	Suhu	Suhu tubuh pasien (℃)	Rasio
X_{11}	Sesak	0 : Pasien tidak mengalami sesak 1 : Pasien mengalami sesak nafas	Nominal
X_{12}	Batuk	0 : Pasien tidak mengalami batuk 1 : Pasien menunjukkan gejala batuk	Nominal
X_{13}	Nadi	Denyut nadi pasien COVID	Rasio
X_{14}	Pilek	0 : Pasien tidak menderita pilek 1 : Pasien menderita pilek	Nominal
X_{15}	Anosmia	0 : Pasien mampu mencium bau 1 : Pasien mengalami kehilangan kemampuan mencium bau	Nominal

C. Struktur Data Penelitian

Berdasarkan variabel penelitian pada Tabel 2, data rekam medis pasien COVID-19 di Rumah Sakit Universitas Airlangga memiliki struktur data pada Tabel 3.

Tabel 3.
Struktur Data Penelitian

Pengamatan ke-	Variabel Respon	Variabel Prediktor					
	Y	X_1	X_2	⋯	X_m	⋯	X_{15}
1	Y_1	$X_{1,1}$	$X_{2,1}$	⋯	⋯	⋯	$X_{15,1}$
2	Y_2	$X_{1,2}$	$X_{2,2}$	⋯	$X_{m,2}$	⋯	$X_{15,2}$
⋮	⋮	⋮	⋮	⋱	⋱	⋱	⋮
n	Y_n	$X_{1,n}$	$X_{2,n}$	⋱	$X_{m,n}$	⋱	$X_{15,n}$
⋮	⋮	⋮	⋮	⋱	⋱	⋱	⋮
668	Y_{668}	$X_{1,668}$	$X_{2,668}$	⋯	$X_{m,668}$	⋯	$X_{15,668}$

D. Langkah-langkah Analisis Data

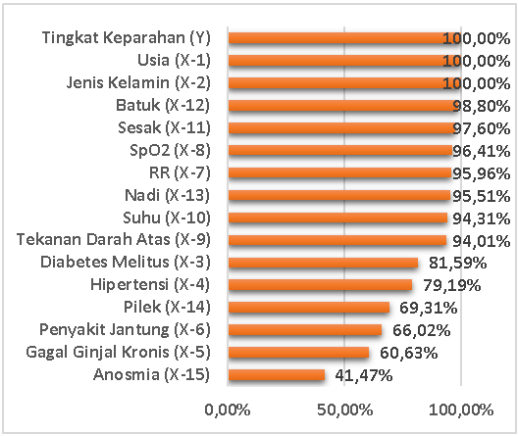
Penelitian ini menggunakan tahapan dalam analisis sebagai berikut.

1. Merangkum data sekunder yaitu rekam medis pasien COVID-19 di RS UNAIR pada periode Mei hingga Oktober 2021 dengan variabel pada Tabel 2.
2. Melakukan uji dependensi dan karakteristik *missing value*.
3. Melakukan imputasi untuk melengkapi data yang hilang dengan metode *k*NN.
4. Melakukan analisis karateristik data tingkat keparahan pasien COVID-19 berdasarkan variabel yang mempengaruhinya.
5. Melakukan penanganan *imbalanced datasets* dengan *Combine Sampling* (SMOTE-Tomek Links).
6. Membentuk model prediktif RF-SVM menggunakan hasil imputasi *k*NN dengan 5-*fold Cross Validation*.
7. Melakukan evaluasi performa model dengan *confusion matrix*, nilai AUC, akurasi, dan F1-*score*.
8. Melakukan interpretasi model terhadap tingkat keparahan pasien COVID-19.

IV. HASIL DAN PEMBAHASAN

A. Penanganan Missing Value

Imputasi *k*NN digunakan dalam penelitian ini untuk mereplikasi data rekam medis yang tidak tercatat oleh tenaga medis. Berikut merupakan karakteristik *missing value* dan tingkat keparahan data sebelum dilakukan imputasi.



Gambar 4. Persentase Data *Non-Missing* Sebelum Imputasi

Gambar 4 menunjukkan karakteristik data sebelum dilakukan imputasi di mana persentase data rekam medis yang tercatat oleh rumah sakit pada data berkisar antara 100% untuk variabel tingkat keparahan, usia, jenis kelamin hingga variabel gejala anosmia pasien yang hanya memiliki 41,47% data yang tercatat. Pada permasalahan *missing value*, perlu diperhatikan jenis data yang hilang pada data tersebut. Perlu diketahui jenis permasalahan data yang hilang sehingga dapat disimpulkan metode penanganan yang sesuai. Selanjutnya dilakukan uji independensi terhadap faktor tingkat keparahan yang ditampilkan dalam Tabel 4.

Tabel 4.
Uji Independensi Prediktor Terhadap Respon

Nama Variabel	Statistik Uji	Df	P-value	Keputusan
Usia (X_1)	84,31	2	0	Tolak H_0
Jenis Kelamin (X_2)	6,218	2	0,045	Tolak H_0
Diabetes Melitus (X_3)	30,167	2	0	Tolak H_0
Hipertensi (X_4)	39,502	2	0	Tolak H_0
Laju Pernafasan (X_7)	111,44	2	0	Tolak H_0
Saturasi Oksigen (X_8)	142,32	2	0	Tolak H_0
Tekanan Darah Atas (X_9)	2,56	2	0,278	Gagal tolak H_0
Suhu (X_{10})	7,67	2	0,022	Tolak H_0
Sesak (X_{11})	90,408	2	0	Tolak H_0
Batuk (X_{12})	2,78	2	0,249	Gagal tolak H_0
Nadi (X_{13})	21,28	2	0	Tolak H_0
Pilek (X_{14})	7,767	2	0,021	Tolak H_0
Anosmia (X_{15})	7,893	2	0,019	Tolak H_0

Menggunakan tingkat signifikansi 5%, tabel diatas menunjukkan bahwa menggunakan tingkat kepercayaan 95% pada data sebelum penanganan imputasi, diperoleh bahwa variabel tekanan darah atas dan gejala batuk tidak berpengaruh signifikan terhadap label dari tingkat keparahan yang berbeda. Sedangkan pada variabel usia, jenis kelamin, riwayat penyakit diabetes, hipertensi, laju pernafasan, saturasi oksigen, suhu, sesak, nadi, pilek, dan anosmia menunjukkan pengaruh yang signifikan terhadap respon. Pada variabel gagal ginjal dan penyakit jantung tidak terdefinisi menggunakan uji *chi-square* dikarenakan ketimpangan yang besar antara pasien yang pernah menderita penyakit tersebut dan tidak sehingga tidak dimasukkan pada tabel hasil uji. Dalam pengecekan tipe *missing value*, digunakan *software* SPSS dalam mendeskripsikan karakteristik pola data yang hilang pada penelitian ini, dihasilkan beberapa karakteristik berikut.

Tabel 5.
Pola *Missing Value* Pada Data

Kasus	Missing (%)	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
212	6,3						S										
213	6,3																S
214	6,3																S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
650	50,0				S	S	S	S	S			S				S	S
651	50,0				S	S	S	S	S	S							S
652	50,0				S	S	S	S					S	S		S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
666	62,5				S	S	S	S	S		S	S			S	S	S
667	62,5				S	S	S	S			S		S	S	S	S	S
668	68,8						S	S	S	S	S	S	S	S	S	S	S

Tabel 5 menunjukkan bahwa terdapat pola data yang hilang di mana persentase data yang hilang lebih dari 50% pada pengamatan 650 hingga 668. Mengindikasikan jumlah *missing value* yang besar, penghapusan 19 pengamatan tersebut dapat dijadikan pertimbangan tersendiri jika terindikasi tipe keacakan data yang hilang adalah *missing completely at random*. Pada tabel di atas juga mengindikasikan pada variabel diabetes, hipertensi, gagal ginjal, penyakit jantung, pilek dan anosmia memiliki lebih banyak data yang hilang dibandingkan variabel lain. Hal ini juga mendukung Gambar 4 di mana 6 variabel tersebut memiliki persentase data yang hilang lebih dari 10%. Selanjutnya dilakukan pengecekan tipe *missing value*, digunakan *software* SPSS dengan uji hipotesis berikut.

H_0 : Data merupakan *missing completely at random* (MCAR)
H_1 : Data bukan merupakan *missing completely at random*
Tingkat kepercayaan = 0,05 (5%)
Statistik uji = *Little's* MCAR hitung = 1390,5
Df = 966
P-value = 0

Daerah kritis = Chi-square tabel = 1039,418

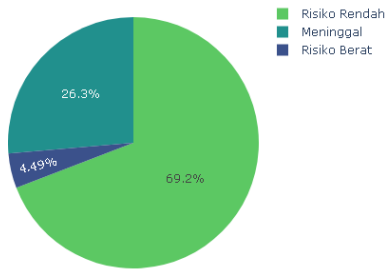
Berdasarkan uji statistik di atas, karena nilai chi-square hitung > daerah kritis dan *p-value* > tingkat kepercayaan (α) maka tolak H_0 sehingga diperoleh kesimpulan bahwa data rekam medis pasien COVID-19 dengan data yang hilang bukan bertipe *missing completely at random* dan disimpulkan tipe data yang hilang ialah *missing at random* atau *missing not at random*. Dengan tipe data yang hilang tersebut, sehingga dapat digunakan imputasi *k*NN dalam mereplikasi data [8].

Setelah penggunaan imputasi *k*NN, data yang hilang telah berhasil ditangani sepenuhnya, di mana data yang baru menunjukkan bahwa karakteristik data pada variabel numerik dan kategorik memiliki persebaran data yang berbeda-beda. Variabel usia pasien memiliki persebaran data dengan nilai antara 0 tahun (bayi baru lahir) hingga 92 tahun dan rata-rata usia pasien adalah 50 tahun serta pasien dengan usia lebih rendah cenderung memiliki tingkat keparahan yang rendah. Variabel laju pernafasan pada pasien memiliki persebaran data dengan rentang 16 hingga 52 nafas per menit dan rata-rata laju pernafasan pasien adalah 24 kali/menit serta pasien dengan kriteria meninggal yang disebabkan kondisi kritis menunjukkan laju pernafasan yang lebih tinggi (24 – 30 /menit) dibandingkan

pasien dengan risiko rendah dan berat (20 – 24 /menit). Variabel saturasi oksigen (SpO2) dalam darah memiliki persebaran data antara 35% hingga 100% dan rata-rata saturasi oksigen pasien ialah 92,8% yang menunjukkan bahwa pasien memiliki kadar oksigen di bawah kriteria normal dengan SpO_2 normal ialah 95% - 100%. Variabel saturasi oksigen juga menunjukkan bahwa pasien dengan tingkat risiko lebih rendah memiliki saturasi oksigen yang lebih tinggi daripada pasien dengan risiko berat dan dinyatakan meninggal. Variabel tekanan darah atas/tekanan darah sistolik adalah tekanan darah saat jantung memompa darah, memiliki persebaran data antara 65 mmHg hingga 225 mmHg dengan rata-rata tekanan darah atas ialah 129 mmHg. Variabel suhu tubuh pasien menunjukkan persebaran data antara 30^0 C hingga 40^0 C dengan rata-rata suhu tubuh ialah $36,7^0$ C. Variabel denyut nadi pasien memiliki persebaran data antara 0 denyut/menit (bpm) yang menunjukkan bahwa pasien meninggal saat pemeriksaan rumah sakit hingga 160 bpm dengan rata-rata denyut nadi pasien ialah 99,5 bpm. Variabel jenis kelamin memiliki rasio antara perempuan dan laki-laki yang seimbang. Variabel riwayat penyakit diabetes pasien menunjukkan bahwa pasien sebagian besar (78%) tidak pernah menderita diabetes melitus, serta terlihat pula bahwa pasien yang berpotensi meninggal memiliki proporsi yang lebih besar saat menderita diabetes. Variabel pasien yang menderita hipertensi menunjukkan bahwa 72% pasien tidak menderita hipertensi serta pasien dengan tingkat keparahan yang lebih tinggi menunjukkan kecenderungan lebih besar mengidap hipertensi. Variabel riwayat penyakit gagal ginjal pasien menunjukkan bahwa sebagian besar pasien (99%) tidak pernah menderita gagal ginjal kronis. Variabel riwayat penyakit jantung juga menunjukkan bahwa 97,6% pasien tidak pernah menderita penyakit jantung. Variabel gejala sesak menunjukkan bahwa terdapat 49% pasien yang menunjukkan gejala ini dengan kecenderungan lebih tinggi terhadap tingkat keparahan yang lebih tinggi. Variabel gejala batuk menunjukkan bahwa sebesar 75% pasien mengidap batuk saat terinfeksi COVID-19. Variabel gejala pilek menunjukkan bahwa sebesar 46% pasien. Variabel gejala anosmia menunjukkan bahwa sebagian besar (89%) pasien tidak kehilangan fungsi indra penciuman.

B. Penanganan Imbalanced Dataset

Karakteristik data yang tidak seimbang ditunjukkan dengan proporsi dari respon yang jauh berbeda antar labelnya, ditunjukkan pada Gambar 5.



Gambar 5. Tingkat Keparahan Pasien COVID-19

Gambar 5 menunjukkan bahwa data memiliki kelas yang tidak seimbang yakni persentase pasien yang berisiko berat sangat sedikit (5%) dan pasien dengan risiko keparahan rendah hingga sedang memiliki persentase yang besar (69%). *Combine sampling* digunakan sebagai penanganan *imbalanced* dataset dengan menggabungkan metode SMOTE yang efektif dalam menghasilkan data sintetis yang mempertahankan karakteristik/pola data awal dan mengurangi bias pada model serta metode Tomek Links yang mengurangi *noise* pada data sehingga metode ini akan menghasilkan dataset baru dengan label kelas yang seimbang. Hasil dari metode ini pada Tabel 6, menunjukkan bahwa penggunaan SMOTE dan Tomek Links efektif dalam menghasilkan 3 kelas respons dengan jumlah yang seimbang.

Tabel 6.
Jumlah Pengamatan Respons Data Setiap Kelas

Label/*Class*	Sebelum Penanganan	Setelah Penanganan
0 = meninggal	176	453
1 = risiko berat	30	456
2 = risiko ringan	462	449

C. Optimasi Parameter

Pemilihan parameter yang optimal sangat diperlukan pada permodelan prediktif, terutama pada model *machine learning* SVM. Terdapat beberapa parameter yang berpengaruh besar terhadap ketepatan prediksi, antara lain parameter *c* yang mengatur *trade-off* antara kesalahan klasifikasi dan margin maksimum, parameter *kernel* yang digunakan untuk memetakan data ke dimensi yang lebih tinggi, parameter *gamma* untuk mengatur seberapa kuat pengaruh titik data dalam pembentukan *hyperplane*, dan parameter *decision function* yaitu jenis pendekatan *one-against-all* dan *one-against-one* pada kasus klasifikasi multi label. Digunakan metode *Grid Search CV* untuk mengoptimasi parameter pada model prediktif. Menggunakan parameter ***c*** = {0,1; 1; 10; 100}, ***gamma*** = {0,1; 1; 10; 100}, ***kernel*** = {linear; polynomial; RBF}, dan ***decision function*** = {*one-against-all* (OVR); *one-against-one* (OVO)} diperoleh hasil optimasi parameter *Grid Search* dengan 5 *cross-validation* menunjukkan kriteria parameter optimal pada model *multiclass* dengan *c* = 10, *gamma* = 1, *kernel* = *Gaussian* RBF, *decision function* = *one-against-all*.

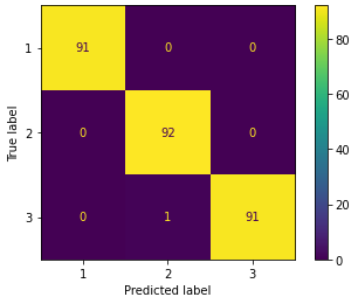
D. Model Prediksi Tingkat Keparahan COVID-19

Model prediktif dalam memperkirakan tingkat keparahan pasien COVID-19 dibangun menggunakan algoritma *hybrid* RF-SVM *multiclass* dengan pendekatan *One-Against-All* (OAA). Pendekatan OAA menggunakan parameter yang telah diuji pada bagian sebelumnya akan digunakan pada permodelan prediktif RF-SVM. Berikut merupakan hasil dari performa model dengan parameter optimal menggunakan *stratified* 5-*cross validation* sebagai pembagi data *training* dan *testing*.

Tabel 7.
Hasil Klasifikasi RF-SVM dengan Imputasi *k*NN

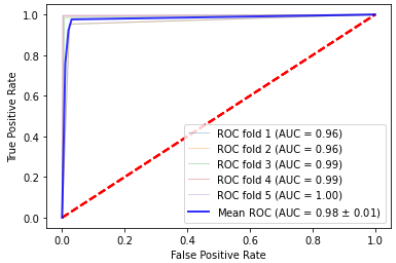
Split/*Fold*	Akurasi (%)	Presisi (%)	*Recall* (%)	F1-*Score* (%)
1	95,29	95,28	95,29	95,27
2	95,29	95,39	95,29	95,30
3	98,55	98,59	98,53	98,54
4	99,27	99,29	99,27	99,27
5	99,64	99,64	99,64	99,64

Berdasarkan nilai performa model di atas, diketahui bahwa pada *split* ke-5, model mampu memprediksi data lebih optimal di mana dihasilkan akurasi dan F1-score sebesar 99,64%. Diperoleh juga pada parameter optimal, yaitu fungsi *kernel* RBF, *c* = 10, dan *gamma* = 1 dihasilkan model OAA dengan kriteria untuk data *testing* memiliki rata-rata akurasi dan F1-score sebesar 99,056% serta performa pada data *training* yaitu rata-rata akurasi dan F1-*score* sebesar 97,6%. Dari kriteria kebaikan model pada data *training* dan *testing* mengindikasikan tidak terdapat perbedaan yang signifikan sehingga dapat disimpulkan bahwa model memiliki generalisasi yang baik serta rendahnya *overfitting* pada model prediktif. Hasil evaluasi performa model *hybrid* RF-SVM *multiclass* dengan menggunakan imputasi *k*NN juga ditampilkan dalam *confusion matrix*. Hasil kebaikan prediksi model pada *split* ke-5 dengan performa tertinggi ditampilkan pada Gambar 6.

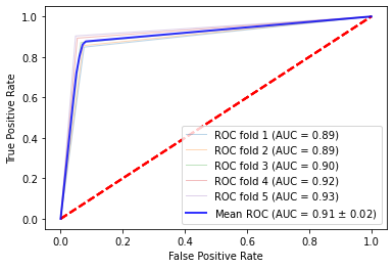


Gambar 6. *Confusion Matrix Hybrid* RF-SVM

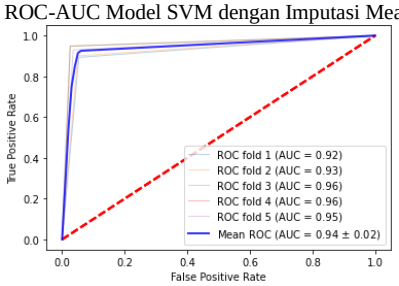
Pada fokus penelitian, analisis juga dilakukan terhadap perlakuan yang berbeda pada prediksi, di mana model diuji menggunakan model SVM *multiclass* dengan imputasi sederhana mean/modus, SVM *multiclass* dengan imputasi *k*NN, serta RF-SVM *multiclass* dengan imputasi mean/modus. Hasil perbandingan kebaikan model antara 4 perlakuan yang berbeda ditunjukkan dengan Gambar 7 hingga Gambar 10.



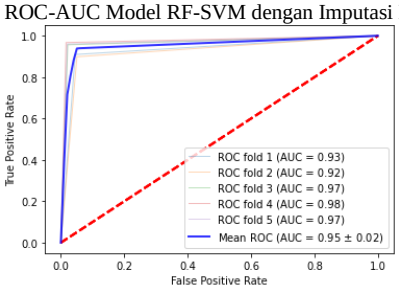
Gambar 7. Kurva ROC-AUC Model RF-SVM dengan Imputasi *k*NN



Gambar 8. Kurva ROC-AUC Model SVM dengan Imputasi Mean/Modus



Gambar 9. Kurva ROC-AUC Model RF-SVM dengan Imputasi Mean/Modus



Gambar 10. Kurva ROC-AUC Model SVM dengan Imputasi *k*NN

Berdasarkan indikator AUC di atas diperoleh hasil klasifikasi pada perlakuan model *hybrid* RF-SVM *multiclass* menggunakan teknik imputasi *k*NN menghasilkan performa yang lebih baik dibandingkan tanpa penggunaan kedua metode tersebut. Hal ini diperlihatkan pada kurva ROC-AUC di mana diperoleh nilai AUC pada tiap model bernilai sebesar 96% hingga 100% pada *split cross validation* serta didapati rata-rata AUC sebesar 98% dengan *standar error* sebesar 1%. Sedangkan pada permodelan tanpa penggunaan metode *hybrid* serta imputasi *k*NN diperoleh nilai AUC berkisar antara 89% hingga 93% dengan rata-rata AUC yaitu 91% dan *standar error* 2%. Berdasarkan hasil performa model di atas, disimpulkan bahwa penggunaan metode imputasi *k*NN lebih baik dalam menghasilkan replika data yang hilang dengan performa yang lebih baik serta model *hybrid* dengan teknik *ensemble* mampu mereduksi efek sehingga diperoleh prediksi yang lebih baik pada data baru. Dibandingkan dengan penelitian terdahulu oleh [13] dimana digunakan studi kasus yang sama sebagai referensi, dengan perlakuan model yang berbeda. Pada penelitian terdahulu digunakan metode SVM *multiclass* tanpa dilakukan imputasi, didapatkan perbendingan berikut.

Tabel 8.
Perbandingan Perlakuan Model dengan Penelitian Terdahulu

Model	Training (%)		Testing (%)		AUC (%)
	Akurasi	F1-*Score*	Akurasi	F1-*Score*	
RF-SVM + Imputasi *k*NN	99,056	99,056	97,608	97,604	98
SVM + Imputasi Mean/Modus	94,054	93,978	87,446	87,182	91
RF-SVM +Imputasi Mean/Modus	93,242	93,152	92,392	92,288	94
SVM + Imputasi *k*NN	100	100	93,834	93,816	95
Penelitian terdahulu	100	100	90,05	90	90,62

Perbandingan diatas menunjukkan bahwa permodelan pada penelitian terdahulu memiliki performa yang tidak begitu bagus dalam memprediksi data *out-sample*. Kesimpulan tersebut diperlihatkan pada Tabel 4.14 dimana pada penelitian terdahulu menghasilkan performa F1-*score* dan akurasi pada data *training* mencapai 100% dan pada data *testing* hanya mencapai 90% dimana hal ini menunjukkan generalisasi yang kurang begitu baik pada data baru. Jika dibandingkan dengan model menggunakan *hybrid* RF-SVM dan imputasi *k*NN terlihat perbedaan performa sebesar 8% dimana hal ini cukup signifikan dalam menentukan penanganan yang sesuai terhadap pasien COVID-19.

V. KESIMPULAN DAN SARAN

Kesimpulan yang diperoleh berdasarkan analisis dan pembahasan adalah sebagai berikut: (1) Karakteristik data pada pasien COVID-19 setelah dilakukan imputasi ialah pasien mayoritas (69%) pengidap COVID-19 termasuk berisiko rendah hingga sedang dengan karakteristik jenis kelamin tidak berpengaruh besar terhadap tingkat keparahan, pasien berusia muda memiliki keparahan lebih rendah, tingkat keparahan pasien COVID-19 akan lebih tinggi terhadap pasien pengidap hipertensi dan diabetes, tingkat keparahan lebih tinggi cenderung menunjukkan gejala laju pernafasan yang lebih tinggi, saturasi oksigen yang lebih rendah dan mengalami sesak nafas. (2) Model prediktif menggunakan *hybrid* RF-SVM memiliki kebaikan model dengan rata-rata F1-*score* mencapai 97,6% dan rata-rata AUC sebesar 98% dengan parameter optimal c = 10, *gamma* = 1, fungsi *kernel* = RBF, dan *decision function = one-against-all*. (3) Perlakuan imputasi *k*NN menunjukkan bahwa metode ini mampu menghasilkan performa prediksi yang lebih baik serta dengan metode *hybrid* RF-SVM menunjukkan bahwa metode ini mampu secara efektif mengurangi *overfitting* dan meningkatkan kemampuan generalisasi model. Berdasarkan analisis yang telah dilakukan pada penelitian ini, diharapkan pihak rumah sakit dapat mengantisipasi perawatan yang tepat dengan lebih cepat menggunakan hasil analisis tersebut di mana sebagian besar pasien terjangkit masih menunjukkan risiko rendah.

DAFTAR PUSTAKA

[1] Breiman, L. (2001). Random Forest. *Machine Learning*(45), 5-32.

[2] Cabitza, F., Campagner, A., Ferrari, D., Resta, C. D., Ceriotti, D., Sabetta, E., . . . Carobene, A. (2021). Development, Evaluation, and Validation of Machine Learning Models for COVID-19 Detection Based on Routine Blood Tests. *Clin Cen Lab Med, II*(59), 421-431.

[3] Cahyo, L. B. (2018). *Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi pada Data Bioinformatika.* Yogyakarta: UII: Tugas Akhir.

[4] Carreras, D. V., Alcaraz, J., & Landete, M. (2023). Comparing Two SVM Models Through Different Metrics Based on the Confusion Matrix. Computers and Operations Research.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321-357.

[6] Daniel, W. W. (2010). Applied Nonparametric Statistic (2 ed.). Michigan: PWS-KENT Pub.

[7] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A Survey on Missing Data in Machine Learning. *Journal of Big Data*.

[8] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis* (Ke7 ed.). Harlow: Pearson Education Limited.

[9] Hairani. (2021). Peningkatan Kinerja Metode SVM Menggunakan Metode KNN Imputasi dan k-Means-SMOTE Untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora. *Jurnal Teknologi Informasi dan Ilmu Komputer* , 713-718.

[10] Hamming, R. W. (1950). Error Detecting and Error Correction Codes. *The Bell System Technical Journal, XX*(2), 147-160.

[11] Karthikeyan, V., & Suja, P. S. (2022). Adaptive Boosted Random Forest-Support Vector Machine based Classification Scheme for Speaker Identification. *Applied Soft Computing*.

[12] Liu, L., Liang, J., Ma, L., Zhang, H., Li, Z., & Liang, S. (2023). Gas Pipeline Flow Prediction Model Based on LSTM with Grid Search Parameter Optimization. *Processes, 11*(1), 63.

[13] Oktaviana, S. M. (2022). *Prediksi TIngkat Keparahan Pasien COVID-19 di Rumah Sakit Universitas Airlangga Surabaya Menggunakan Combine Sampling Support Vector Machine.* Surabaya: ITS.

[14] Oktaviani, I. D., & Putrada, A. G. (2022). KNN Imputation to Missing Values of Regression-based Rain Duration Prediction on BMKG Data. *Jurnal Informatics - Telecommunication - Electronics*, 249-254.

[15] Rao, T., & Rajinikanth, T. V. (2014). A Hybrid Random Forest based Support Vector Machine Classification Supplemented by Boosting. *Global Journal of Computer Science and Technology, I*(14), 43-54.

[16] Saguna, G. N. (2019). *Klasifikasi Tweet Terhadap Layanan Customer Care XL Axiata dengan Metode Naive Bayes dan Random Forest.* Surabaya: Tugas Akhir ITS.

[17] SATGAS. (2021, Oktober 31). Analisis Data COVID-19 Indonesia Update per 31 Oktober 2021. *Analisis Data COVID-19 Indonesia*, hal. 1-179.

[18] Scholkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge: MIT.

[19] Steinwart, I., & Christmann, A. (2008). *Support Vector Machines.* New York: Springer Science+Business Media.

[20] Sudriani, Y., Setiawan, F. A., & Hamid, A. (2020). Comparison of kNN and Iterative Imputation Approach for Missing Data Value of Online Water Quality Monitoring System in Lake Maninjau. *Jurnal Online Informatika*.

[21] Tomek, I. (1998). Two Modification of CNN. *IEEE Transactions on System, Man, and Communication*, 769-772.

[22] WHO. (2021). *Living Guidance for Clinical Management of COVID-19.* Geneve: WHO Press.

[23] Yilmaz, A. E., & Demirhan, H. (2023). Weighted Kappa Measures for Ordinal Multi-class Classification Performance. *Applied Soft Computing*.

[24] Zhang, R., Xiao, Q., Zhu, S., Lin, H., & Tang, M. (2021). Using Different Machine Learning Models to Classify Patient Into Mild and Severe Cases of COVID-19 Based on Multivariate Blood Testing. *Journal of Medical Virology*, 357-365

[25] Zhang, S. (2012). Nearest Neighbor Selection for Iteratively kNN Imputation. *The Journal of Systems and Software*, 2541-2552.

[26] Zhang, X., & Liu, C. (2022). Model Averaging Prediction by K-Fold Cross-Validation. *Journal of Econometrics*.