Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2478~2485
e-ISSN: 2548-6861

Prediction of Tuberculosis Treatment Outcomes in Indonesia Using Support Vector Machine and Random Forest

Dian Sugianto [1]*, Joko Triloka [2]*
* Fakultas Ilmu Komputer, Institut Informatika Dan Bisnis Darmajaya Lampung
dian.2421211033p@mail.darmajaya.ac.id [1], joko.triloka@darmajaya.ac.id [2]

Article Info

Article history:

Received 2025-06-30
Revised 2025-07-10
Accepted 2025-07-19

Keyword:

Tuberculosis,
Support Vector Machine,
Random Forest,
Recovery Prediction,
Machine Learning.

ABSTRACT

Tuberculosis (TB) remains a global health challenge, particularly in developing countries such as Indonesia, which ranks third worldwide in the number of TB cases. This study aims to evaluate the performance of Support Vector Machine (SVM) and Random Forest (RF) algorithms in predicting TB patient recovery rates based on clinical data obtained from healthcare facilities in Indonesia. Evaluation results indicate that the model achieved very high precision scores (100%) for the "Deceased," "Transferred," and "Default" categories; however, these findings require critical interpretation due to the likely class imbalance in those categories. In contrast, for the "Recovered" and "Completed" categories—where data instances were more numerous—the model exhibited lower precision and recall values (below 90%), reflecting challenges in accurately predicting majority classes. These results suggest that despite seemingly high numerical performance, model predictions can be biased if class distribution is not appropriately considered. The main contribution of this research lies in providing a comparative analysis of two widely used machine learning algorithms in predicting TB recovery outcomes, while emphasizing the importance of addressing data imbalance issues in clinical predictive modeling. The findings provide a practical basis for integrating predictive algorithms into clinical workflows, enabling more accurate monitoring of patient recovery and timely adjustments of TB treatment plans in Indonesia.

I. PENDAHULUAN

Tuberkulosis (TB) adalah penyakit menular yang masih menjadi masalah kesehatan global, termasuk di Indonesia. Berdasarkan data dari Organisasi Kesehatan Dunia (*WHO*), TB merupakan salah satu dari sepuluh penyebab utama kematian di dunia. Indonesia berada di peringkat ketiga dalam jumlah kasus TB global, dengan estimasi 824.000 kasus baru setiap tahun pada 2021. Tingginya angka kasus ini menunjukkan perlunya peningkatan dalam strategi penanganan dan prediksi kesembuhan pasien TB di Indonesia[1][2].

Upaya penanggulangan TB di Indonesia sebenarnya telah dilakukan, bahkan sejak tahun 1995, namun, secara umum belum menunjukkan keberhasilan yang signifikan, bahkan pada tahun 2006-2022 angka kesembuhan TB di Indonesia cenderung mengalami tren penurunan. Pada tahun 2022, rata-rata persentase kesembuhan penyakit TB di Indonesia sebesar 46,5% [3].

Meskipun kemajuan dalam pengobatan TB telah dicapai, prediksi kesembuhan pasien TB masih menghadapi tantangan signifikan. Data klinis pasien, seperti riwayat pengobatan, komorbiditas, respons imunologi, dan faktor demografis, sering kali bersifat heterogen, tidak lengkap, dan non-linear. Pendekatan prediktif tradisional, seperti regresi logistik atau analisis statistik deskriptif, memiliki keterbatasan dalam menangani kompleksitas data tersebut, terutama ketika melibatkan interaksi antar-faktor yang sulit dimodelkan secara linear [4] [5].

Kesembuhan pasien TB dipengaruhi oleh berbagai faktor seperti jenis pengobatan, kondisi kesehatan umum, dan kepatuhan terhadap terapi. Tantangan utama yang dihadapi tenaga medis adalah dalam merancang strategi pengobatan

yang efektif dengan memanfaatkan data yang kompleks dan beragam [6].

Di sisi lain, perkembangan Machine Learning (ML) menawarkan peluang untuk meningkatkan akurasi prediksi kesembuhan TB. Meskipun telah ada penelitian yang menggunakan metode seperti decision tree C4.5 dalam prediksi kesembuhan pasien TB [7], model ML tersebut cenderung mengorbankan interpretabilitas demi performa (black box), padahal interpretabilitas sangat penting bagi tenaga medis dalam memahami faktor dominan yang memengaruhi kesembuhan pasien dan menghindari kesalahan diagnosis. Fokus pada interpretabilitas untuk TB masih menjadi kesenjangan penelitian karena belum banyak dieksplorasi, terutama dalam konteks pengembangan model yang seimbang antara akurasi dan kemudahan interpretasi.

Solusi berbasis ML untuk prediksi TB saat ini didominasi oleh pendekatan tunggal seperti penggunaan Support Vector Machine (SVM) atau Random Forest (RF) yang memiliki kelemahan spesifik yaitu SVM sulit diinterpretasi meskipun akurat untuk data tidak linier. Sedangkan RF cenderung *overfitting* jika hiper parameternya tidak dioptimalkan. Algoritma SVM dan RF telah digunakan secara luas dalam bidang kesehatan untuk tugas klasifikasi dan prediksi, termasuk memprediksi kesembuhan pasien TB. SVM berfokus pada pencarian hiperplane optimal untuk membedakan kelas data, sementara RF menggunakan metode ensemble dengan menggabungkan prediksi dari berbagai pohon keputusan untuk meningkatkan stabilitas dan akurasi hasil prediksi. Kedua algoritma ini unggul dalam menangani data dengan dimensi tinggi dan kompleks, serta mampu mengatasi masalah data hilang (missing values) yang sering dijumpai dalam data medis [8-15].

Meskipun telah ada penelitian yang menggunakan metode seperti SVM dan RF ataupun metode lain seperti regresi logistik dan decision tree dalam prediksi kesembuhan pasien TB, penerapan algoritma SVM dan RF masih relatif jarang dilakukan di Indonesia. Keterbatasan masing-masing metode machine learning dalam memprediksi kesembuhan pasien TB mendorong perlunya evaluasi komparatif untuk mengidentifikasi pendekatan teroptimal. SVM dikenal unggul dalam menangani data berdimensi tinggi dengan overfitting minimal, sementara Random Forest memiliki kemampuan untuk menangkap interaksi kompleks antar variabel klinis. Namun, belum ada konsensus tentang metode mana yang lebih efektif bila diterapkan pada konteks data TB di Indonesia, yang kerap memiliki karakteristik unik seperti keterbatasan sampel, ketidaklengkapan rekam medis, dan variasi faktor epidemiologi.

Disisi lain, algoritma Support Vector Machine (SVM) dan Random Forest (RF) telah banyak digunakan dalam prediksi penyakit, penelitian yang secara langsung membandingkan kedua algoritma tersebut dalam konteks prediksi kesembuhan pasien TB di Indonesia masih sangat terbatas. Sebagian besar studi sebelumnya hanya menggunakan satu pendekatan algoritmik tanpa mempertimbangkan keunikan distribusi data lokal yang cenderung tidak seimbang, tidak lengkap, dan heterogen. Selain itu, belum banyak penelitian yang secara eksplisit membahas dampak ketidakseimbangan data terhadap performa model prediksi, terutama pada kategori dengan jumlah sampel yang rendah seperti “Meninggal” atau “Default”. Hal ini menciptakan kesenjangan penelitian dalam hal evaluasi menyeluruh terhadap keandalan model pembelajaran mesin di sektor kesehatan masyarakat.

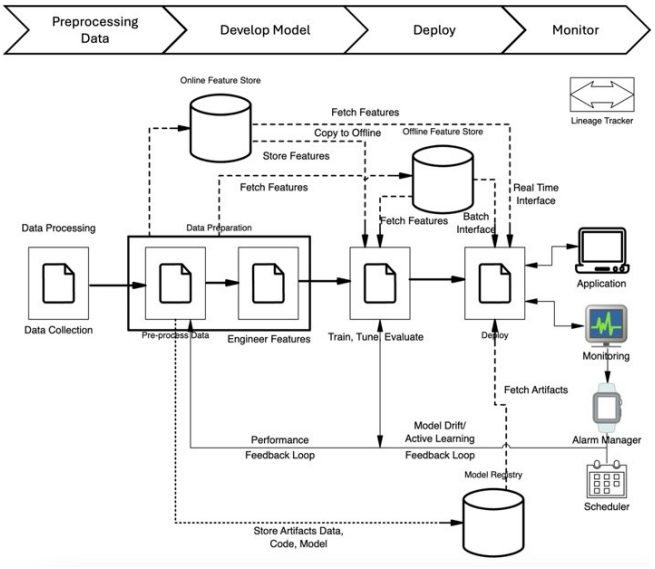
Penelitian ini menawarkan kontribusi kebaruan dengan melakukan evaluasi komparatif antara SVM dan RF menggunakan data klinis lokal, sekaligus menyoroti pentingnya penanganan ketidakseimbangan data dalam membangun model prediktif yang andal. Selain itu, penelitian ini menunjukkan bahwa nilai precision yang tinggi pada kategori minoritas dapat bersifat menyesatkan jika tidak dikaji bersama konteks distribusi data, sehingga memberikan wawasan baru dalam pemodelan klasifikasi medis berbasis machine learni

Penelitian ini bertujuan membandingkan kinerja SVM dan RF dalam memprediksi kesembuhan pasien TB menggunakan dataset lokal, untuk menentukan pendekatan yang paling akurat dan stabil. Hasilnya diharapkan dapat menjadi dasar rekomendasi bagi tenaga medis dalam memilih model prediksi yang sesuai dengan kondisi fasilitas kesehatan di Indonesia, sekaligus mendukung pengambilan keputusan klinis yang lebih berbasis bukti [16][17].

II. METODE

Metode penelitian terdiri dari beberapa tahapan, secara umum meliputi tahap pengumpulan data, *preprocessing* data, *develop* model yang meliputi pelatihan model, evaluasi, dan analisis hasil, tahap implementasi (*deploy*) dan monitoring dengan desain arsitektur disjaikan pada Gambar 1.

Pendekatan arsitektur digunakan pada dataset kecil atau masalah yang dapat diselesaikan dengan model prediktif berbasis aturan yang jelas, dalam hal ini menggunakan data klinis pasien penderita TB.



Gambar.1 Arsitektur Model Prediksi SVM dan Random Forest

A. Pengumpulan Data

Penelitian diawali dengan pengumpulan dataset klinis TB meliputi gejala pasien, hasil laboratorium (tes dahak dan X-ray), riwayat medis, serta data faktor demografi dari rumah sakit yang dipastikan sudah memenuhi kriteria kelengkapan dan relevansi untuk prediksi akurat.

B. Preprocessing Data

Data mentah dibersihkan dengan menangani missing values (imputasi), *outlier*, dan ketidakseimbangan kelas (SMOTE), kemudian dilakukan normalisasi (Min-Max) dan encoding fitur kategorikal untuk memastikan kesiapan data sebelum pemodelan.

Dalam menangani ketidakseimbangan kelas yang ditemukan dalam dataset, pendekatan SMOTE (Synthetic Minority Over-sampling Technique) digunakan selama tahap preprocessing. SMOTE bekerja dengan mensintesis data baru untuk kelas minoritas berdasarkan kemiripan lokal antar sampel, sehingga meningkatkan representasi kategori yang jarang seperti "Meninggal", "Pindah", dan "Default". Teknik ini dipilih karena lebih efektif dibandingkan metode undersampling yang berpotensi menyebabkan hilangnya informasi penting dari kelas mayoritas. Selain SMOTE, penyesuaian class weight pada algoritma juga dipertimbangkan, namun tidak digunakan dalam eksperimen utama ini karena fokus utama penelitian adalah membandingkan performa dua algoritma (SVM dan RF) dalam kondisi input data yang telah diseimbangkan secara eksternal. Evaluasi performa model dilakukan setelah proses oversampling untuk memastikan bahwa peningkatan akurasi pada kategori minoritas bukan sekadar hasil bias data, melainkan representasi valid dari kemampuan model dalam menangani distribusi yang lebih seimbang.

C. Develop Model (Pelatihan, Evaluasi dan Analisis Hasil)

Menggunakan algoritma Support Vector Machine (SVM) dan Random Forest untuk melatih model dengan data-data pelatihan. Kemudian menilai kinerja model menggunakan data uji berdasarkan berbagai metrik evaluasi, seperti akurasi, sensitivitas, spesifisitas, dan AUC (Area Under Curve) serta menganalisis hasil prediksi untuk menentukan faktor yang paling berpengaruh terhadap tingkat kesembuhan pasien

D. Implementasi (Deploy)

Tahapan ini adalah proses mengimplementasikan model prediksi ke dalam sistem nyata sehingga dapat digunakan untuk memprediksi tingkat kesembuhan pasien secara langsung.

E. Monitoring (Pemantauan Kinerja Model)

Memastikan bahwa model tetap bekerja sesuai dengan ekspektasi setelah diterapkan dalam lingkungan nyata

F. Lokasi dan Waktu Penelitian

Penelitian ini menggunakan data klinis yang diperoleh dari rumah sakit dan klinik penanganan pasien tuberkulosis (TB) di Bandar Lampung. Data bersumber dari Sistem Pencatatan dan Pelaporan SITK (Sistem Informasi Tuberkulosis Komunitas) serta dikumpulkan atas izin lembaga Inisiatif Lampung Sehat dan Electronic Medical Record (EMR) dari puskesmas, klinik, dan rumah sakit, sesuai kebutuhan penelitian. Pengumpulan data berlangsung selama tiga bulan, sedangkan proses analisis dan pengujian model dilakukan dua bulan setelahnya.

G. Data dan Sumber Data

Data yang digunakan dalam penelitian ini meliputi beberapa data. Data Demografi, terdiri dari 635 field/record data yang mencakup informasi demografi. Data ini meliputi variabel-variabel utama seperti jenis kelamin, usia, pekerjaan, dan wilayah. Analisis terhadap data ini bertujuan untuk memahami distribusi dan karakteristik populasi yang menjadi objek penelitian. Diagram data demografi ditunjukkan pada Gambar 2.

Status pengobatan pasien dalam dataset ini terbagi ke dalam lima kategori utama, yaitu:

1. Sembuh (Recovered) sebanyak 266 pasien (41.89%),
2. Lengkap (Completed) sebanyak 346 pasien (54.49%),
3. Meninggal (Deceased) sebanyak 12 pasien (1.89%),
4. Pindah (Transferred) sebanyak 8 pasien (1.26%),
5. Default (Putus Berobat) sebanyak 3 pasien (0.47%).

Distribusi ini menunjukkan adanya ketimpangan kelas yang signifikan, di mana dua kategori utama ("Sembuh" dan "Lengkap") mendominasi lebih dari 96% data, sedangkan tiga kategori lainnya memiliki representasi yang sangat kecil. Kondisi ini menjadi tantangan tersendiri dalam pelatihan model prediktif, karena model cenderung bias terhadap kelas mayoritas jika ketidakseimbangan tidak ditangani secara eksplisit.

Untuk proses pelatihan dan evaluasi, data dibagi menjadi dua bagian menggunakan skema train-test split sebesar 80:20, sehingga:

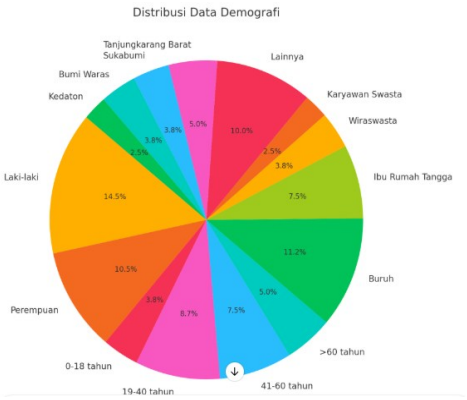
a) 508 data digunakan sebagai data pelatihan (training set), dan
b) 127 data sebagai data pengujian (testing set).

Proses pemisahan data dilakukan secara stratified random split untuk menjaga proporsi distribusi kelas dalam kedua subset, sehingga evaluasi model menjadi lebih representatif dan adil. Ketidakseimbangan ini kemudian ditangani dalam tahap preprocessing menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) guna meningkatkan performa model pada kategori minoritas.

Data Pengobatan: Riwayat pengobatan pasien, seperti jenis obat yang diberikan, dosis, durasi pengobatan, serta informasi kepatuhan pasien terhadap regimen pengobatan. Data ini dikumpulkan dari catatan medis pasien TB yang telah menerima pengobatan selama minimal 6 bulan dan telah mendapatkan hasil evaluasi kesembuhan. Data-data tersebut berasal dari sumber catatan medis rumah sakit atau fasilitas kesehatan yang merawat pasien, serta laporan evaluasi pengobatan yang dilakukan oleh tenaga medis. Data ini

penting untuk menilai efektivitas pengobatan yang diterima pasien dan memantau adanya potensi resistensi obat atau komplikasi yang dapat mempengaruhi prognosis pasien [18].

1. Rekam Medis Elektronik (EMR) dari fasilitas kesehatan seperti rumah sakit dan puskesmas.
2. Survei Pasien yang dilakukan oleh petugas kesehatan selama proses pengobatan



Gambar 2. Demografi Usia, Jenis Kelamin, Pekerjaan dan Wilayah

H. Metode Pengumpulan Data

Metode pengumpulan data dalam penelitian ini meliputi [19]. Dokumentasi sebagai pengambilan data dari rekam medis elektronik yang berisi informasi mengenai diagnosis dan hasil pengobatan pasien TB. Data ini diakses dengan izin dari lembaga terkait dan melalui proses etis yang ketat untuk menjaga kerahasiaan informasi pasien.

Survei, dilakukan kepada pasien yang telah menyelesaikan pengobatan untuk mendapatkan informasi tambahan mengenai faktor-faktor yang mungkin memengaruhi hasil pengobatan, seperti kepatuhan terhadap pengobatan dan kondisi sosial-ekonomi.

Wawancara, dilakukan dengan petugas kesehatan (dokter dan perawat) yang bertanggung jawab atas pengobatan pasien TB untuk memperoleh informasi lebih rinci mengenai proses perawatan.

I. Metode Pengolahan Data

Data yang telah dikumpulkan akan diproses dan diolah menggunakan metode [7]. Langkah awal dalam pengolahan data dalam konteks machine learning adalah melakukan preprocessing untuk memastikan bahwa data berada dalam kondisi yang siap digunakan oleh model. Preprocessing merupakan tahap krusial karena kualitas data akan sangat memengaruhi performa model yang dibangun Salah satu aspek penting dalam preprocessing adalah penanganan data yang hilang. Jika ditemukan nilai-nilai yang kosong dalam dataset, maka akan dilakukan proses imputasi, yaitu pengisian nilai kosong tersebut menggunakan metode yang sesuai, seperti nilai rata-rata, median, modus, atau teknik lain yang lebih kompleks, tergantung pada karakteristik dan pola data yang tersedia. Dengan mengatasi missing data secara sistematis, model akan lebih stabil dan hasil prediksi menjadi lebih andal. Selain itu, karena sebagian data yang digunakan dalam machine learning sering kali bersifat kategorikal—seperti data jenis kelamin, status pekerjaan, atau wilayah tempat tinggal—maka diperlukan proses konversi ke bentuk numerik. Proses ini dilakukan melalui teknik encoding, dan salah satu metode yang digunakan adalah label encoding, di mana setiap kategori diberi label berupa angka. Langkah ini penting agar model dapat memproses fitur-fitur tersebut secara matematis. Langkah berikutnya dalam preprocessing adalah normalisasi data numerik. Normalisasi bertujuan untuk menyesuaikan skala dari setiap fitur agar berada dalam rentang yang sebanding. Hal ini dilakukan untuk mencegah fitur-fitur dengan nilai besar mendominasi fitur lainnya dalam proses pembelajaran model. Dengan demikian, semua fitur memiliki kontribusi yang seimbang dalam proses training dan prediksi. Melalui tahapan preprocessing yang sistematis, kualitas data akan meningkat dan model machine learning yang dibangun akan memiliki kinerja yang lebih optimal dan dapat diandalkan dalam menghasilkan prediksi

Pembagian dataset. Setelah data diproses, data akan dibagi menjadi dua set: data pelatihan (training data) dan data uji (testing data) dengan perbandingan 80:20. Data pelatihan akan digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa model.

Dalam penelitian ini, algoritma Support Vector Machine (SVM) dan Random Forest digunakan untuk memprediksi tingkat kesembuhan pasien. SVM bekerja dengan mencari hyperplane terbaik yang dapat memisahkan pasien sembuh dan tidak sembuh, menggunakan kernel yang disesuaikan dengan pola data. Sementara itu, Random Forest membentuk sejumlah pohon keputusan dari subset acak data pelatihan, lalu menghasilkan prediksi akhir melalui voting mayoritas dari seluruh pohon. Kedua model kemudian dievaluasi untuk dibandingkan performanya dalam klasifikasi tingkat kesembuhan pasien.

Setelah proses pelatihan selesai, performa kedua model dievaluasi menggunakan data uji untuk mengetahui sejauh mana kemampuan model dalam melakukan prediksi yang akurat. Evaluasi dilakukan menggunakan beberapa metrik utama, yaitu akurasi, sensitivitas, spesifisitas, dan AUC. Akurasi mengukur proporsi prediksi yang benar terhadap seluruh data uji. Sensitivitas mengukur seberapa baik model dalam mengenali pasien yang benar-benar sembuh, sedangkan spesifisitas menunjukkan kemampuan model dalam mengidentifikasi pasien yang tidak sembuh. Selain itu, AUC digunakan untuk menilai kemampuan model dalam membedakan antara dua kelas, yaitu sembuh dan tidak sembuh. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki performa yang sangat baik dalam klasifikasi [20].

J. Metode Analisis Data

Setelah melakukan pengolahan dan evaluasi model, analisis data dilakukan untuk mengidentifikasi faktor-faktor yang paling memengaruhi kesembuhan pasien TB. Dengan

menggunakan Random Forest, fitur yang dominan (*feature importance*) akan dianalisis untuk menentukan variabel-variabel klinis yang memiliki dampak terbesar terhadap hasil pengobatan. Hasil dari model ini kemudian akan dibandingkan untuk menentukan algoritma yang memberikan performa terbaik [21].

Analisis ini mencakup analisis deskriptif yakni menyajikan gambaran umum data pasien, seperti distribusi usia, jenis kelamin, status nutrisi, serta hasil pengobatan yang dicapai.

Analisis perbandingan yakni membandingkan hasil prediksi dari kedua model SVM dan Random Forest untuk menentukan algoritma yang lebih akurat dalam memprediksi tingkat kesembuhan pasien TB.

Analisis faktor penting yakni mengidentifikasi variabel klinis yang paling berpengaruh dalam prediksi tingkat kesembuhan, berdasarkan hasil model Random Forest.

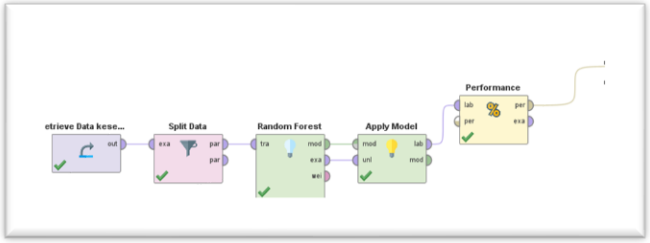
Tabel 1 merangkum fitur-fitur yang digunakan dalam pemodelan prediksi kesembuhan pasien TB berdasarkan data klinis, survei, dan rekam medis.

III. HASIL DAN PEMBAHASAN

Penerapan algoritma *Random Forest* yang telah dioptimalkan diproses menggunakan perangkat RapidMiner sebagaimana ditunjukkan pada Gambar 3.

TABEL 1 DAFTAR ATRIBUT (FITUR) INPUT YANG DIGUNAKAN DALAM MODEL PREDIKSI

No	Nama Fitur	Jenis Data	Sumber Data	Keterangan
1	Usia	Numerik	Rekam Medis / EMR	Usia pasien saat memulai pengobatan TB
2	Jenis Kelamin	Kategorikal	Rekam Medis / EMR	Laki-laki atau perempuan
3	Pekerjaan	Kategorikal	Survei Pasien	Status pekerjaan (buruh, PNS, wiraswasta, dll.)
4	Wilayah Tempat Tinggal	Kategorikal	Survei Pasien / EMR	Lokasi domisili pasien (kecamatan/kota)
5	Gejala Klinis	Kategorikal	Wawancara Petugas / EMR	Misalnya: batuk berdarah, demam, berat badan turun, dll.
6	Hasil Tes Dahak	Kategorikal/Numerik	Hasil Laboratorium	Hasil tes sputum, termasuk kuantitas kuman (jika tersedia)
7	Hasil Rontgen (X-Ray)	Kategorikal	Hasil Radiologi	Indikasi kerusakan paru atau tanda-tanda TB aktif
8	Jenis Obat yang Diberikan	Kategorikal	Catatan Pengobatan	Regimen pengobatan (misalnya HRZE, HR, dll.)
9	Dosis dan Durasi Obat	Numerik	Catatan Pengobatan	Jumlah dan lama konsumsi masing-masing obat
10	Kepatuhan Minum Obat	Kategorikal	Survei Petugas / EMR	Tinggi, sedang, rendah; berdasarkan pengawasan langsung atau laporan
11	Riwayat Pengobatan Sebelumnya	Kategorikal	Rekam Medis	Pernah atau belum pernah mendapat pengobatan TB sebelumnya
12	Status Gizi (jika tersedia)	Kategorikal/Numerik	Survei / Rekam Medis	Indeks massa tubuh (IMT) atau kategori status gizi pasien (baik, kurang)



Gambar 3. Proses *Random Forest* dengan Rapid Miner

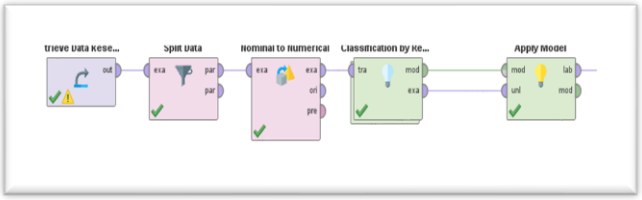
Selanjutnya hasil proses dari RapidMiner dapat dilihat pada Tabel 2 yang menunjukkan performa model dalam mengklasifikasikan status pasien TB berdasarkan lima kategori utama: "Sembuh," "Lengkap," "Meninggal," "Pindah," dan "Default."

TABEL 2
HASIL PROSES RANDOM FOREST

true Sembuh	true Lengkap	true Meninggal	true Pindah	true Default	class Precision
178	46	6	1	0	77.06%
34	234	1	4	1	85.40%
0	0	3	0	0	100.00%
0	0	0	2	0	100.00%
0	0	0	0	1	100.00%
83.96%	83.57%	30.00%	28.57%	50.00%	

Tabel 2 menunjukkan hasil evaluasi model machine learning dalam memprediksi status kesembuhan pasien TB, dengan lima kategori: "Sembuh," "Lengkap," "Meninggal," "Pindah," dan "Default." Model memiliki precision tinggi untuk kategori "Meninggal," "Pindah," dan "Default" (100%) tetapi lebih rendah pada "Sembuh" (77.06%) dan "Lengkap" (85.40%). Dari segi recall, kategori "Sembuh" dan "Lengkap" memiliki kinerja yang cukup baik (masing-masing 83.96% dan 83.57%), sedangkan recall untuk kategori "Meninggal," "Pindah," dan "Default" lebih rendah (30%, 28.57%, dan 50%). Hal ini menunjukkan bahwa meskipun model mampu memprediksi beberapa kategori dengan baik, terdapat tantangan dalam mengidentifikasi beberapa status lain secara akurat.

Berikutnya adalah penerapan algoritma SVM yang diproses menggunakan perangkat RapidMiner sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Proses SVM dengan Rapid Miner

Hasil proses dari RapidMiner menunjukkan performa model dalam mengklasifikasikan status pasien TB berdasarkan lima kategori utama: "Sembuh," "Lengkap," "Meninggal," "Pindah," dan "Default" sebagaimana disajikan pada Tabel 3.

Confusion matrix pada Tabel 3 menunjukkan performa model yang sangat baik dalam mengklasifikasikan status pasien TB ke dalam lima kategori: "Sembuh," "Lengkap," "Meninggal," "Pindah," dan "Default."

TABEL 3
HASIL PROSES SUPPORT VECTOR MACHINE

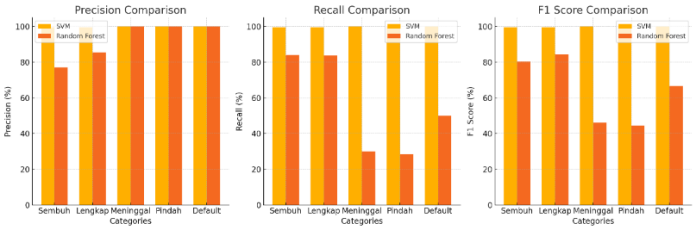
true Sembuh	true Lengkap	true Meninggal	true Pindah	true Default	class precision
211	1	0	0	0	99.53%
1	279	0	0	0	99.64%
0	0	10	0	0	100.00%
0	0	0	7	0	100.00%
0	0	0	0	2	100.00%
99.53 %	99.64 %	100.00 %	100.00 %	100.00 %	

Model mencapai *precision* dan *recall* yang tinggi, dengan sebagian besar kategori memiliki nilai 99.5% atau lebih. Prediksi "Sembuh" dan "Lengkap" masing-masing memiliki *precision* 99.53% dan 99.64%, sementara kategori "Meninggal," "Pindah," dan "Default" memiliki precision sempurna 100%, menunjukkan tidak ada kesalahan prediksi untuk kategori tersebut. *Recall* juga sangat tinggi di semua kategori, dengan nilai 100% untuk "Meninggal," "Pindah," dan "Default," serta hampir sempurna untuk "Sembuh" dan "Lengkap," menunjukkan kemampuan model yang akurat dan efektif dalam mengidentifikasi status pasien dengan benar.

Dalam analisis lanjutan, kinerja SVM sangat menonjol karena mampu menangani data dengan lebih baik dan menghasilkan prediksi yang sangat akurat di hampir semua kategori. Kelebihan SVM terletak pada kemampuannya untuk membedakan dengan sangat baik antara kelas-kelas yang berbeda, bahkan pada kategori yang sulit seperti "Meninggal" dan "Pindah". Dengan precision dan recall mendekati atau sama dengan 100% pada semua kategori, SVM menunjukkan bahwa model ini tidak hanya mampu membuat prediksi yang tepat, tetapi juga dapat menangkap semua kategori tanpa kehilangan banyak informasi. Selain itu SVM memiliki margin yang baik dalam menangani kesalahan klasifikasi dan sangat optimal dalam tugas ini.

Sebaliknya, Random Forest meskipun memiliki performa yang lumayan baik pada kategori "Sembuh" dan "Lengkap", mengalami kesulitan dalam mendeteksi kasus yang lebih jarang seperti "Meninggal" dan "Pindah", yang terlihat dari rendahnya *recall* pada kategori tersebut. Kelemahan ini disebabkan oleh karakteristik model *Random Forest* yang kesulitan dalam menangani distribusi data yang tidak seimbang atau jumlah sampel yang lebih kecil dalam kategori tertentu. Meski *precision*-nya tinggi, rendahnya *recall* pada beberapa kategori membuat model ini kurang optimal dibandingkan SVM, terutama dalam skenario di mana deteksi yang tepat dari semua kategori sangat penting. Secara grafik perbandingan kedua algoritma ditunjukkan pada Gambar 5.

Grafik pada Gambar 5 menunjukkan perbandingan antara SVM dan Random Forest dalam hal Precision, Recall, dan F1 Score untuk setiap kategori (Sembuh, Lengkap, Meninggal, Pindah, Default).



Gambar 5. Grafik Perbandingan SVM dan Random Forest

Secara global perbandingan ditunjukkan pada Tabel 4 dimana SVM memiliki keunggulan signifikan dalam semua metrik, baik dari sisi akurasi maupun kesetaraan performa di seluruh kelas. Nilai macro F1-score yang sangat tinggi menunjukkan bahwa SVM mampu menangani kelas mayoritas dan minoritas secara seimbang, menjadikannya model yang lebih stabil dan andal dalam konteks klasifikasi multikategori seperti prediksi kesembuhan pasien TB.

TABEL 4
PERBANDINGAN METRIK GLOBAL

Metrik Evaluasi	Random Forest	SVM
Accuracy	81.80%	99.61%
Macro F1-score	64.42%	99.83%
Micro F1-score	81.80%	99.61%

Berdasarkan hasil pada Tabel 4, dapat disimpulkan bahwa model Support Vector Machine (SVM) menunjukkan performa yang jauh lebih baik dibandingkan dengan Random Forest pada seluruh metrik evaluasi. SVM mencapai akurasi sebesar 99.61%, macro F1-score sebesar 99.83%, dan micro F1-score sebesar 99.61%, sedangkan Random Forest hanya memperoleh akurasi 81.80%, macro F1-score 64.42%, dan micro F1-score 81.80%.

Berdasarkan analisis grafik, terlihat bahwa SVM memiliki akurasi 99.51%, jauh lebih tinggi dibandingkan dengan Random Forest yang hanya mencapai 81.96%. Dari sisi precision, SVM menunjukkan nilai yang sangat tinggi dan

hampir sempurna pada semua kategori, sementara Random Forest memiliki precision yang lebih rendah, terutama pada kategori “Sembuh” dan “Lengkap”. Pada metrik recall, SVM juga menunjukkan kinerja yang konsisten dan sangat tinggi di setiap kategori, sedangkan Random Forest mengalami penurunan recall yang cukup signifikan, khususnya pada kategori “Meninggal” dan “Pindah”.

Selain itu, hasil F1-score memperlihatkan bahwa SVM memiliki nilai yang tinggi dan stabil, menandakan adanya keseimbangan yang baik antara precision dan recall. Sebaliknya, Random Forest mengalami penurunan F1-score yang cukup besar pada beberapa kategori seperti “Meninggal” dan “Pindah”, yang menunjukkan bahwa model ini masih mengalami kesulitan dalam mendeteksi seluruh kelas secara akurat. Secara keseluruhan, SVM terbukti menjadi model dengan performa terbaik dan paling konsisten dalam mengklasifikasikan data dibandingkan Random Forest.

IV. KESIMPULAN

Dari hasil analisis perbandingan kinerja antara Support Vector Machine (SVM) dan Random Forest berdasarkan berbagai metrik evaluasi seperti precision, recall, dan F1 score, dapat disimpulkan bahwa SVM menunjukkan performa yang lebih superior dalam menangani tugas klasifikasi untuk semua kategori. SVM memiliki nilai precision, recall, dan F1 score yang sangat tinggi dan stabil, terutama pada kategori yang lebih menantang seperti "Meninggal" dan "Pindah". Hal ini menunjukkan bahwa SVM tidak hanya akurat dalam membuat prediksi, tetapi juga mampu mendeteksi seluruh kelas dengan baik tanpa kehilangan informasi penting.

Sementara itu, Random Forest meskipun menunjukkan performa yang cukup baik pada beberapa kategori seperti "Sembuh" dan "Lengkap", mengalami penurunan yang signifikan dalam recall dan F1 score pada kategori "Meninggal" dan "Pindah". Temuan ini mengindikasikan bahwa Random Forest memiliki keterbatasan dalam mendeteksi kategori dengan jumlah sampel yang lebih sedikit atau distribusi yang tidak seimbang.

Meskipun Random Forest memiliki nilai akurasi keseluruhan yang lebih tinggi (99,51%) dibandingkan SVM (81,96%), hal ini disebabkan oleh dominasi kelas mayoritas dalam dataset. Ketika mempertimbangkan precision, recall, dan F1-score pada semua kategori secara seimbang terutama pada kelas minoritas seperti "Meninggal" dan "Pindah", SVM terbukti lebih stabil dan akurat. Oleh karena itu, secara keseluruhan SVM lebih unggul dalam konteks prediksi multikategori yang tidak seimbang.

Secara keseluruhan, SVM lebih unggul dalam menjaga keseimbangan antara precision dan recall, khususnya dalam skenario di mana deteksi menyeluruh terhadap semua kategori sangat krusial. Dengan demikian, SVM dapat dianggap lebih efektif untuk diterapkan dalam tugas klasifikasi yang bersifat kompleks dan kritikal.

DAFTAR PUSTAKA

[1] A. C. Kurniawan, A. Salam, “Seleksi Fitur Information Gain untuk Optimasi Klasifikasi Penyakit Tuberkulosis”, *J. Media Inform. Budidarma*, vol. 8, no. 1, p. 70, 2024, doi: 10.30865/mib.v8i1.7122.

[2] R. Melati N, T. Waluyo Purboyo, M. Kalista, “Prediksi Penderita Tuberkulosis Menggunakan Algoritma Support Vector Regression (SVR)”, *e-Proceeding Eng.*, vol. 10, no. 1, pp. 736–741, 2023.

[3] A. N. Dita, D. M. Nadia, S. Toha, Suliyanto, "Modeling the Percentage of Tuberculosis Cure in Indonesia Using a Multivariate Adaptive Regression Spline Approach", *Inferensi*, vol.7 no.2. pp. 91-97, 2024.

[4] S. Maren, E et al. “Developing prediction models for clinical use using logistic regression: an overview.” *Journal of thoracic disease* vol. 11, Suppl 4: S574-S584, 2019. doi:10.21037/jtd.2019.01.25

[5] I. Madakkatel, A. Zhou, M. D. McDonnell et al., "Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study", *Sci Rep 11*, 22997, 2021. https://doi.org/10.1038/s41598-021-02476-9

[6] S. Deepti, B. Pronaya, A. Verma et al., "Explainable AI for Healthcare 5.0: Opportunities and Challenges", *IEEE Access*. 1-30, 2022. 10.1109/ACCESS.2022.3197671.

[7] R. Rusdah and B. A. Bregastantyo, “Model Prognosis Masa Pengobatan Pasien Tuberkulosis Dengan Metode C4.5,” *J. Teknol. Inf. dan Ilmu Komput*., vol. 10, no. 6, pp. 1197–1204, 2023, doi: 10.25126/jtiik.1067393.

[8] K. Deny, W. Mochamad, P. Lise et al. “Deteksi dan Prediksi Cerdas Penyakit Paru-Paru dengan Algoritma Random Forest”. Indonesian Journal Computer Science. 3. 51-56. 2024. 10.31294/ijcs.v3i1.6071

[9] A. Christian, H. Hariyanto, A. Yani, S. Sumanto, “Analisis Machine Learning Untuk Prediksi Penyakit Paru-Paru Menggunakan Random Forest”. Journal of Innovation and Future Technology (IFTECH), 7(1), 122-133.2025. https://doi.org/10.47080/iftech.v7i1.3906

[10] L. Rangga A. T, D. Dahlan, “Optimalisasi Fitur Dengan Forward Selection Pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma Klasifikasi Random Forest, JATI, Vol. 8 No. 5, 2024. https://doi.org/10.36040/jati.v8i5.11064

[11] E. Priyono, “Prediction of Tuberculosis Patients With Machine Learning Algorithms", J. Ilm. Penelit. dan Pembelajaran Inform., vol.9, no.4, 2024. https://doi.org/10.29100/jipi.v9i4.5486

[12] B. S. C. Putra, I. Tahyudin, B. A. Kusuma, K. N. Isnaini, “Efektivitas Algoritma Random Forest, XGBoost , dan Logistic Regression dalam Prediksi Penyakit Paru-paru”, Techno.Com, 23(4), 909–922, 2024.

[13] M. Andani, J. Triloka, S. Y. Irianto, H. W. Nugroho, “Comparison of K-Nearest Neighbor, Naive Bayes, Random Forest Algorithms for Obesity Prediction”, Sinkron: Jurnal Dan Penelitian Teknik Informatika, 9(1), 502-510, 2025. https://doi.org/10.33395/sinkron.v9i1.14478

[14] T. M. Prasetyo, A. Amrullah, S. Syahrir, B. N. Sari, “Implementasi Algoritma SVM (Support Vector Machine) Dalam Klasifikasi Penyakit Paru-Paru Berdasarkan Fitur Pola Bentuk”, Jurnal Teknologi Informasi, 6(1), 1–6, 2022

[15] A. Sandika, F. R. Ramadhan, I. N. Iman, J. Jihad, “Optimasi Prediksi Penyakit Paru-Paru dan Kanker Paru melalui Integrasi Algoritma Random Forest”, Buletin Ilmiah Ilmu Komputer dan Multimedia (BIIKMA), vol. 2, no. 3, 585–591, 2024.

[16] S. S. SH, R. R, J. Partha Sarathy, G. Kishore, S. S and B. Jegajothi, "Ensemble-based Deep Learning Framework for Tuberculosis Detection in Radiographs," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1363-1371, doi: 10.1109/ICEARS64219.2025.10940231.

[17] V. Balakrishnan, G. Ramanathan, S. Zhou,. et al. Optimized support vector regression predicting treatment duration among tuberculosis patients in Malaysia. Multimed Tools Appl 83, 11831–11844 (2024). https://doi.org/10.1007/s11042-023-16028-y

[18] N. Nurfadilla, M. Afdal, I. Permana, and Z. Zarnelly, “Comparison

of Data Mining Algorithm for Clustering Patient Data Human Infectious Diseases," *J. Tek. Inform.*, vol. 4, no. 5, pp. 1127–1134, 2023, doi: 10.52436/1.jutif.2023.4.5.983.

[19] P. N. Isnaeni, H. Rakhmawati, F. A. Tyas, S. Muhammadiyah, and P. Brebes, "Perbandingan Algoritma Naïve Bayes Dan C4.5 Pada Klasifikasi Penyakit Tuberculosis," J. Informatics Comput., vol. 3, no. 1, pp. 41–48, 2024.

[20] H. Saiyar, "Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes," *Jurikom)*, vol. 5, no. 5, pp. 498–502, 2018, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom%7CPage%7C498

[21] E. Mutiara, "Algoritma Klasifikasi Naive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Tuberculosis (Tb)," Swabumi, vol. 8, no. 1, pp. 46–58, 2020, doi: 10.31294/swabumi.v8i1.7668